Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa : kriittisiä havaintoja

Nelimarkka, Matti

2019

Nelimarkka , M 2019 , ' Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa : kriittisiä havaintoja ' , Politiikka : Valtiotieteellisen yhdistyksen julkaisu , Vuosikerta. 61 , Nro 1 , Sivut 6-33 . < https://journal.fi/politiikka/article/view/79629 >

http://hdl.handle.net/10138/300868

publishedVersion

Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: kriittisiä havaintoja

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

MATTI NELIMARKKA

ABSTRAKTI Aihemallinnus mahdollistaa laajojen tekstiaineistojen automaattisen ryhmittelyn käyttämällä ohjaamatonta koneoppimista. Kiinnostus aihemallinnusta kohtaan on kasvanut ja sen soveltaminen on lisääntynyt yhteiskuntatieteellisessä tutkimuksessa. Aihemallinnus sekä muut ohjaamattoman koneoppimisen menetelmät kuitenkin vaativat tutkijoita tekemään valintoja: tutkijat joutuvat esimerkiksi päättämään mitä koneoppimismenetelmää käytetään, miten sitä käytetään ja miten aineistoa esikäsitellään, Lisäksi on kyettävä tulkitsemaan ohjaamattoman koneoppimisen kautta syntyneet tulokset. Aihemallinnuksessa eräs valinta koskee aiheiden määrää, josta on käyty aktiivisesti keskustelua niin koneoppimisen kuin laskennallisen yhteiskuntatieteen yhteisöissä. Artikkelin esittelemä käyttäjäkoe osoittaa, että yhteiskuntatieteissä suosittu, tulkinnallisuutta korostava lähestymistapa aihemäärän valintaan on epävarma. Artikkelin empiirinen esimerkki osoittaa, että aihemäärän valinta vaikuttaa aihemallinnuksesta syntyviin tulkintoihin. Tämän pohjalta artikkeli suosittaa, että (i) parametrien valinnassa käytettäisiin tilastollisia menetelmiä. Lisäksi suositellaan, että (ii) aihemallinnuksen tulokset sidotaan yhteiskuntatieteelliseen kirjallisuuteen käyttämällä teoreettista viitekehystä tulkinnan apuna tai aihemallinnusta käytetään joko menetelmällisesti trianguloiden tai *grounded theory* -lähtöisesti. Lisäksi artikkelissa suositellaan, että (iii) tutkimusprosessin avoimuuteen kiinnitetään huomiota sekä (iv) laskennallisten menetelmien soveltajat seuraavat kriittisen algoritmitutkimuksen kehitystä.

JOHDANTO

Laadullisten sekä määrällisten menetelmien lisäksi yhteiskuntatieteissä on viime aikoina sovellettu laskennallisia *(computational)* menetelmiä. Laskennalliset menetelmät sekä digitaalisen murroksen tuomat massa-aineistot (*big data*) ovat yhdessä mahdollistaneet laskennallisen yhteiskuntatieteen (*computational social science*) käyttöönoton (Cioffi-Revilla 2010; Lazer ym. 2009). Näiden menetelmien käyttö on kasvanut merkittävästi 2000-luvulla (Savage 2013). Menetelmien on myös sanottu haastavan yhteiskuntatieteen tietokäsitystä sekä menetelmällisiä lähtökohtia (Boyd ja Crawford 2012; Kitchin 2014).

Laskennallisia menetelmiä voidaan soveltaa myös tekstiaineistojen analyysiin (esim. Grimmer ja Stewart 2013). Tekstiaineistojen tilastolliseen ryhmittelyyn perustuva aihemallinnus (*topic modelling*) on herättänyt laajasti mielenkiintoa yhteiskuntatieteilijöiden piirissä (esim. Levy ja Franklin 2014; Laaksonen ja Nelimarkka 2018; Purhonen ja Toikka 2016; Ylä-Anttila ym. 2018). Menetelmä on kehitetty noin viisitoista vuotta sitten ja ensimmäinen laaja-alaiseen käyttöön tarkoitettu tieteellinen artikkeli siitä on kirjoitettu vuonna 2012 (Blei 2012; Blei ym. 2003). Uutuutensa takia menetelmälle ja sen soveltamiselle ei ole vakiintuneita tapoja. Vaikka artikkelit korostavat tarvetta validoida laskennallisten menetelmien tuloksia myös tekstianalyysissä, ne eivät kerro kuinka validointi tulisi tehdä (vrt. Grimmer ja Stewart 2013).

Artikkelissa käsitellään aihemallinnukseen liittyviä menetelmällisiä haasteita. Aihemallinnus tuottaa aina tutkijan päättämän määrän ryhmiä aineistosta (aiheita) ja näyttää miten tietyt sanat ja tekstit kuuluvat näihin aiheisiin. Wallach ym. (2009) argumentoivat, että aihemallinnuksen mallin parametrien[1] valinta vaikuttaa aihemallin tuottamiin aiheryhmiin ja muihin aihemallinnuksen tuloksiin. Onkin siis odotettavissa, että myös aihemäärillä (k) tuotetut aihemallit ja niiden tulokset ovat erilaisia. Perinteinen yhteiskuntatieteellinen lähestymistapa aihemäärän valitsemiseksi on ollut kokeilla muutamia erilaisia aihemääriä ja valita niiden perusteella tutkimuskysymyksen kannalta valaisevin aihemäärä (esim. Levy ja Franklin 2014; Purhonen ja Toikka 2016). Tietojenkäsittelytieteessä on taas korostettu tilastollisia mittareita ja niiden soveltamista aiheiden määrän valinnassa (Griffiths ja Steyvers 2004; Wallach ym. 2009). On kuitenkin epäselvää, miten eri tutkijat valitsevat aiheita ja mikä on aihevalinnan merkitys empiirisissä tuloksissa. Tässä artikkelissa pyritään käsittelemään näitä haasteita aihemallinnuksen sekä muiden ohjaamattomien koneoppimismenetelmien kautta.

Artikkelissa käydään ensin läpi tekstiaineiston laskennallista analyysiä yleisesti sekä aihemallinnusta prosessina. Tämän jälkeen artikkeli esittää kaksi osatutkimusta, joiden yhteydessä esitellään itsenäinen kirjallisuuskatsaus sekä tulosten tulkinta keskustelun muodossa. Ensimmäisessä osatutkimuksessa käsitellään aihemäärän valinnan haasteita tuomalla esille toisaalta tutkijoiden eroja tässä valinnassa sekä toisaalta tutkijoiden sekä tilastollisten lähestymistapojen eroja. Toisessa osatutkimuksessa käsitellään tarkemmin aihemäärän valinnan tärkeyttä empiirisessä tutkimusprosessissa. Tätä tarkastellaan soveltamalla ensimmäisen osatutkimuksen aihemääriä empiiriseen tutkimuskysymykseen Suomen puoluekentän kehittymisestä. Artikkeli päättyy johtopäätöksiin siitä, mitä kaksi tutkimuskysymystä kertovat aihemallinnuksen sekä laajemmin ohjaamattomien koneoppimismenetelmien soveltamisesta yhteiskuntatieteissä. Tällöin keskustelussa otetaan kantaa laskennallisten menetelmien kriittisen tutkimuksen puolesta (vrt. esimerkiksi Savage 2013).

TEKSTI LASKENNALLISENA DATANA

Tekstianalyysissä käytettävät laskennalliset menetelmät voidaan pääpiirteissään jakaa kolmeen ryhmään: sanastopohjaiseen, ohjattuun sekä ohjaamattomaan koneoppimiseen (esim. Grimmer ja Stewart 2013). Tämän ryhmittelyn sijaan yhteiskuntatieteissä tutumpi lähestymistapa voi olla jako ”a priori -skeemaan perustuviin menetelmiin” sekä ”aineistolähtöisiin menetelmiin” (Purhonen ja Toikka 2016). Molemmille on vastineet niin perinteisissä yhteiskuntatieteellisissä menetelmissä sekä uusissa laskennallisissa menetelmissä. A priori-skeemoja vastaa erilaisten luokittelukehikkojen soveltaminen: esimerkiksi sisällön luokittelu koodikirjaa käyttäen on tyypillinen a priori -skeema. Aineistolähtöisistä menetelmistä *grounded theory* lienee tunnetuin esimerkki: aineistoon tutustuminen synnyttää siihen jonkun ryhmittelyn.

Näille menetelmille voidaan myös löytää laskennalliset vastineet. *Ohjattu koneoppiminen* vastaa a priori -skeemoja korostavia lähestymistapoja. Siinä käytetään olemassa olevaa luokiteltua aineistoa (mitä usein kutsutaan opetusaineistoksi) ja etsitään siitä laskennallisesti piirteitä – useimmiten sanoja – jotka mahdollisimman hyvin selittävät kuulumista luokkaan. Esimerkiksi Yhdysvalloissa republikaaniedustajat voidaan erottaa noin 60% tarkkuudella demokraattiedustajista tarkastelemalla heidän puheenvuoroja edustajanhuoneessa (Yu ym. 2008). Ohjatun koneoppimisen etuna on tarkkuuden mahdollisimman täsmällinen arviointi: käytössä on sekä opetusaineiston tunnettu luokitus että koneoppimisen kautta laskettu luokitus aineistolle. Näitä kahta mittausta käyttämällä on helppo tarkastella kuinka mallinnusmenetelmät ja näiden erilaiset parametrit vaikuttavat tarkkuuteen. Tarkkuutta voidaan mitata monella tavalla. Koneoppimistutkijoiden parissa yleisiä mittareita ovat ulkoinen tarkkuus (*accuracy*), saanti (*recall*) ja sisäinen tarkkuus (*precision*), kun taas yhteiskuntatieteilijöiden keskuudessa kyseessä on luokittelijoiden välisen eron arviointia, johon käytetään esimerkiksi Cronbachin α tai Cohenin κ -mittareita. Ohjatun koneoppimisen luotettavuuden arvioinnissa sille ilmoitetun tarkkuuden arviointi on keskeistä.

Aineistolähtöisten menetelmien laskennallinen vastine ovat *ohjaamattomat koneoppimismenetelmät*. Ohjaamattomissa koneoppimismenetelmissä ei tehdä ennakkoon oletuksia aineiston luokittelukriteereistä. Sen sijaan aineistolle etsitään laskennallisesti ryhmiä (esimerkiksi aihemallinnus tai *k*-means-menetelmä) tai lainalaisuuksia (esimerkiksi assosiaatiosäännöt, ks. Jurek ja Scime 2014). Ohjaamattomat menetelmät ovat samankaltaisia faktorianalyysin kanssa: molemmat lähestymistavat ovat aineistolähtöisiä.

Vertailu faktorianalyysiin on hyödyllistä, koska se on useille yhteiskuntatieteilijöille tuttu menetelmä. Lisäksi eksploratiivinen faktorianalyysi on ollut käytössä niin kauan, että menetelmää on jo tarkasteltu kriittisesti (esim. Fabrigar ym. 1999; Bandalos ja Boehm-Kaufman 2010; Russell 2002). Psykologian aikakausilehdissä julkaistuista eksploratiivista faktorianalyysiä soveltavista artikkeleista jopa viidenneksessä analyysin tulokset on raportoitu epätarkasti ja menetelmää on käytetty virheellisesti (Fabrigar ym. 1999). Virhelähteitä ovat niin otoskoon pienuus, liian suuri faktorien määrä suhteessa aineiston kokoon, faktorien määrän valinnan perustelut sekä faktorianalyysissä käytetyn menetelmän valinnat (ks. Bandalos ja Boehm-Kaufman 2010, 79–83; Fabrigar ym. 1999; Russell 2002). Artikkelin fokuksessa olevien ohjaamattomien menetelmien kriittinen tarkastelu on tarpeen, koska ohjaamattomaan oppimiseen perustuvat menetelmät ovat jo näyttäneet haastavuutensa.

AIHEMALLINNUS

Aihemallinnus ryhmittelee tekstiä (dokumentteja ja niissä olevia sanoja) ja löytää sieltä piilossa olevia rakenteita tai 'aiheita'. Matemaattisesti aihemallit perustuvat todennäköisyysjakaumiin, joiden avulla tarkastellaan dokumenttien kuulumista aiheeseen sekä sanojen kuulumista aiheisiin. Aihemallinnusta on käytetty esimerkiksi tutkittaessa etujärjestöjen ja kansalaisten kommentteja lakiluonnoksiin (Levy ja Franklin 2014), viestinnän painopisteitä viestintävälineiden välillä (Laaksonen ja Nelimarkka 2018; Nelimarkka ym. arvioitavana), presidenttien puheiden sisällön analyysiin (Purhonen ja Toikka 2016) sekä ilmastonmuutosta koskevien kehysten analyysissä (Ylä-Anttila ym. 2018). Aihemallinnuksessa on kolme vaihetta (kuva 1): esikäsittely, analyysi, sekä tulkinta.

Esikäsittely { Kielen esikäsittely; Aineiston puhdistus

Analyysi { Algoritmin parametrien valinta (mm. aihemäärä); Aihemallin sovittaminen

Tulkinta { Tulosten tulkinta; Validointi

Kuva 1: Aineiston esikäsittely

Aineiston esikäsittely

Koska aihemallinnus perustuu sanojen esiintymisien laskentaan, on välttämätöntä muuttaa sanat niiden perusmuotoon. Tällöin eri sija- ja taivutusmuodoissa esiintyvät sanat muokataan samaan muotoon: esimerkiksi sanat 'kissat', 'kissoja' ja 'kissoille' kaikki viittaavat perusmuotoiseen asiaan 'kissa'. Kielitieteessä on kaksi lähestymistapaa perusmuotoistamiseen. Stemmauksessa sanat katkaistaan kieliopillisten sääntöjen pohjalta katkaistuun sanamuotoon. Lemmaus taas etsii sanalle perusmuotoista ilmaisua sanakirjojen sekä kielellisen analyysin kautta. Näillä kahdella tavalla on eroja lopputuloksiin, kuten taulukossa 1 esitetään. Erityisesti stemmauksessa sanat voivat olla usein vaikeatulkintaisia. Sekä stemmaukseen että lemmaukseen löytyy monia valmiita ratkaisuja. Stemmauksessa *Natural Language Toolkit* (NLTK)[2] on yleisesti käytetty ratkaisu. Lemmaus taas vaatii kieltä ymmärtävän morfologisen jäsentimen. Suomeksi tällainen on käytettävissä esimerkiksi Kielipankin kautta[3].

	Esimerkki 1	**Esimerkki 2**
Alkuperäinen	Kissalla on pitkät viikset	Eilen oli vaalit
Stemmaus	kis on pitk viiks	eile oli vaali
Lemmaus	kissa olla pitkä viiksi	eilen olla vaalia

Taulukko 1: Stemmauksen ja lemmauksen eroja

Kielen esikäsittelyn lisäksi on välttämätöntä puhdistaa aineistoa jatkokäsittelyyn. Esimerkiksi teksti muutetaan pieneen kirjainkokoon ja välimerkit sekä numerot poistetaan. Lisäksi analyysistä on usein hyödyllistä poistaa yleisiä sanoja *(stopwords)*, kuten konjunktioita 'ja', 'tai' ja 'myös', sekä yleisiä verbejä, kuten 'olla'. Samalla tavoin tässä vaiheessa voidaan poistaa sanoja, jotka eivät joko selkeytä tulkintaa tai ovat niin yleisiä, että ne jopa häiritsevät sitä. Toisaalta aineistoa voi puhdistaa laajemminkin, esimerkiksi poistamalla kaikki partitiivit tai erisnimet.

Tällä hetkellä ei ole olemassa selkeää ohjetta aineiston puhdistamiseen. Sen sijaan aineiston puhdistamisessa voidaan tehdä perustellusti hyvinkin erilaisia päätöksiä. Siksi onkin välttämätöntä dokumentoida tehdyt valinnat selkeästi (vrt. Denny ja Spirling 2018). Aihemallinnuksen (ja ohjaamattomien koneoppimismenetelmien) osalta menetelmäkeskustelu on kuitenkin vielä lapsen kengissä. Esimerkiksi perusmuotoistamista on pidetty kriittisenä perinteisissä menetelmäkuvauksissa, mutta tuorein kirjallisuus on haastanut tämän esikäsittelyvaiheen merkityksen. Kuten Schofield ja Mimno (2016) huomauttavat, stemmaus ja lemmaus voivat heikentää aihemallinnuksen laatua englannin kielellä. Toisaalta erilaisia vaihtoehtoja aineistojen puhdistamiseen on valtavasti ja niillä voidaan parantaa tai heikentää ohjaamattoman koneoppimisen tuloksia (Denny ja Spirling 2018). Vaikka puhdistamiseen liittyvä menetelmäkeskustelu on vilkastunut, valitettavasti sitä käydään lähinnä englannin kielen osalta. Esimerkiksi Schofieldinin ja Mimnon (2016) havainnot eivät ole yleistettävissä suomeen: englannissa persoonapäätteitä ja sijapäätteitä ei käytetä samalla tavalla kuin suomessa. Siksi suomenkielisen aineiston käsittelyssä on vielä syytä noudattaa perinteisiä ohjeita aineiston puhdistamisesta: aineisto kannattaa lemmata ja poistaa yleisiä sekä harvinaisia sanoja ennen aihemallinnusta.

Aineiston analyysi aihemallinnuksella

Aihemallinnus laskee jokaiselle tekstin analyysiyksikölle (eli dokumentille) aiheiden jakauman kyseisessä dokumentissa sekä edelleen jokaiselle sanalle jakauman aiheisiin kuulumisesta. Jokaiselle dokumentille ja jokaiselle sanalle siis lasketaan kuuluvuus kaikkiin aiheisiin. Kukin näistä vaihtelee välillä 0–1.

Aihemallinnusalgoritmi

Dokumenttien ja sanojen jakaumat aiheille lasketaan todennäköisyyspohjaisesti. Tällä hetkellä yleinen lähestymistapa aihemallinnukseen on LDA-menetelmä (*Latent Dirichlet Allocation*; Blei ym. 2003; Blei 2012). Dirichlet-jakauma on siis tilastollinen jakauma, kuten yhteiskuntatietelijöille tutummat normaali- ja χ^2-jakaumat. Todennäköisyyspohjainen malli käyttää kolmea hyperparametriä: alkuarvaus aiheiden jakautumisesta dokumenteille (α), alkuarvaus aiheiden jakautumisesta sanoille (β) sekä aiheiden määrä (k). Käytännössä parametrit α ja β vaikuttavat siihen, kuinka herkästi aihemallinnus tulkitsee aiheita esiintyvän eri dokumenteissa ja sanoissa.

Aihemallinnus on laskennallisesti vaativa lähestymistapa: sekä dokumenttien että sanojen jakaumia päivitetään aineistoa läpikäymällä (esim. Blei ym. 2003). Käyttäen Dirichlet-jakaumia jokaiselle sanalle ja dokumentille arvioidaan mahdollinen jakauma niiden kuulumisesta tiettyyn aiheeseen (sanakohtainen jakauma θ ja dokumenttipohjainen jakauma ϕ pohjautuvat α, β arvoihin). Näitä jakaumia parannetaan käymällä läpi jokainen aineiston dokumentti ja sana.

Sekä sanoille että dokumenteille lasketaan Bayes-päättelyllä jatkuvasti uusia θ- sekä ϕ-jakaumia. Toisin sanoen, jokaisen dokumentin jokaisen sanan perustella θ- sekä ϕ-jakaumia päivitetään sanan ja dokumentin muodostaman yhdistelmän luoman uuden havainnon avulla. Prosessia jatketaan, kunnes jokainen dokumentti ja sana on käyty läpi. Näin kiinnostavat lopulliset arvot θ sekä ϕ perustuvat koko aineistoon. Lopulta ne mittaavat sanojen esiintymistä yhdessä tietyissä aiheissa sekä näiden aiheiden esiintymistä dokumenteissa.

Aiheiden määrän valinta

Parametrien α ja β lisäksi aihemallinnuksessa tulee valita aiheiden määrä (k). Kyseessä on haastava vaihe tutkimusprosessissa; esimerkiksi Greene ym. (2014) nostavat aiheiden määrän valinnan keskeiseksi haasteeksi aihemallinnuksessa. Yhteiskuntatieteilijät ovat yleisesti tarkastelleet muutamia eri aihemääriä ja valinneet näistä selkeiten tulkittavan (esim. Purhonen ja Toikka 2016; Levy ja Franklin 2014). Esimerkiksi Levy ja Franklin (2014) perustelevat, että tutkimusprosessissa keskeinen vaihe on tulkita aihemallinnuksen tuloksia. Tämän takia he valitsevat heidän mielestään selkeimpiä aiheita tuottavan aihemäärän. Toisaalta, on olemassa myös tilastollisia mittarisuureita, joiden pohjalta löydettyjen aiheiden soveltuvuutta voidaan arvioida. Silloin tutkijat voivat sovittaa aihemallin usealle eri aihemäärälle ja käyttää tilastollisia suureita päättämään, mikä on paras aihemäärä (esim. Griffiths ja Steyvers 2004; Wallach ym. 2009). Kuten *a priori* -skeemojen arvioinnissa, aihemallinnuksen sopivuuden mittaamisessa on olemassa useita mittareita, kuten perpleksiteetti ja suurimman uskottavuuden harmonisen keskiarvon malli. Perpleksiteetti mittaa mallin sopivuutta aineistoon laskemalla, kuinka usein tuotettu malli loisi alkuperäisen aineiston mukaisia dokumentteja. Tällöin siis pienemmät perpleksiteettiarvot kuvaavat tilannetta, jossa malli on selkeämpi (Blei 2012; Blei ym. 2003). Suurimman uskottavuuden harmonisen keskiarvon malli sen sijaan perustuu aihemallinnuksen logaritmisen uskottavuudem (*loglikelihood*-arvon) tulkintaan. Tämä mittaa mallin ja aineiston välistä sopivuutta. Harmoninen keskiarvo parantaa arvoa verrattuna yksittäiseen logaritmisen uskottavuuden arvoon (Griffiths ja Steyvers 2004; Wallach ym. 2009). Menetelmää on sovellettu useissa lähteissä, koska se on yksinkertainen toteuttaa ja varsin nopea (esim. Griffiths ja Steyvers 2004).

Tietojenkäsittelytieteilijät ovat esittäneet myös tarkempia mittareita aihemallinnuksen tarkkuuden arvioimiseksi. Esimerkiksi Wallach ym. (2009) huomioivat, että Chib-estimaattori (*Chib-style estimator*) ja vasemmalta-oikealle menetelmä (*"left-to-right" algorithm*) voivat tuottaa tarkempia tuloksia. Molemmat menetelmät perustuvat aihemallinnusjakauman arviointiin suhteessa muihin estimoituihin aihemallinnusjakaumiin ja niiden välisen muutoksen tarkasteluun. Näiden käsittely tarkemmin ei ole vielä kuitenkaan tarpeen, koska ne puuttuvat useista valmiista aihemallinnuskirjastoista. Nämä esimerkit kuitenkin osoittavat aihemallinnuksen olevan vielä kehittyvä ala, jonka käytännöt mukautuvat ja muuttuvat. Tämän takia soveltajien on välttämätöntä seurata menetelmäkehityksestä käytävää keskustelua aktiivisesti.

Vaikkakin laskennalliset mittarit selkeyttävät aiheiden määrän valintaa, eivät ne tietenkään ole ongelmattomia. Chang ym. (2009) osoittavat, ettei aihemallin muodostamat ryhmät ole aina täysin selkeitä. Käyttäjäkokeessaan he pyysivät osallistujia jatkamaan aihemallin sanalistaa viiden sanan jälkeen ja arvioimaan tuloksia suhteessa aihemallin oikeisiin sanoihin. Heidän tuloksensa näyttivät, etteivät osallistujat pystyneet tulkitsemaan tilastollisesti mitattuna parhainta

mallia. He päättelivät, että tilastollisen mittarin kannalta parhaan aihemallin aiheet eivät muodostaneet selkeitä kokonaisuuksia[4].

Aihemallinnuksen tulosten tulkinta

Viimeinen vaihe aihemallinnuksessa on aihemallin tulosten tulkinta sekä validointi. Aihemallinnusprosessin päätteenä tarkastellaan sanalistoja aiheittain ja sanaryhmän perusteella aihe nimetään mielekkäästi (ks. liitetaulukko 1). Sanalistojen kautta muodostuvien aiheiden on näytetty tuottavan ihmisille mielekkäimpiä aihekokonaisuuksia (Aletras ym. 2017). Nimeämisprosessi on Purhosen ja Toikan (2016) mukaan samankaltainen kuin faktorianalyysissä. Nimeämiselle ei ole yksikäsitteistä sääntöä – lopputulos on tutkijan tulkinta sanojen ja aiheiden merkityksistä. Usein ne perustuvat aiheiden yleisten tai kuvaavien sanojen käyttöön ja niiden tulkintaan.

Sanalistojen tarkastelun lisäksi tulosten validointi on tekstianalyysissä välttämätöntä (esim. Grimmer ja Stewart 2013). Vaikkakin validoinnin merkitystä korostetaan osana tutkimusprosessia, siihen ei tällä hetkellä tarjota kovinkaan selkeää mallia. Myöskin kirjallisuus validoinnista on varsin heikkoa. Myös tässä artikkelissa pääpaino on aihemallinnuksen suorittamisessa, eikä tuloksia pyritä tarkemmin validoimaan. Artikkelissa kuitenkin esitetään kolme toisistaan poikkeavaa tapaa validoinnin tekemiseen. Koska menetelmän käyttö yhteiskuntatieteissä kehittyy jatkuvasti, voi myös yleisesti hyväksytty validointimenetelmä kehittyä myöhemmin.

Yksinkertaisin tapa on tulkita aihemallinnuksen aiheita koodikirjana aineistolle. Osa aineistosta koodattaisiin uudelleen manuaalisesti käyttäen tätä koodikirjaa. Tämän jälkeen käytettäisiin perinteisiä välineitä tutkijoiden välisen luotettavuuden arviointiin. Lähestymistavan hyvänä puolena on yksinkertaisuus sekä tuttuus kvalitatiivisia menetelmiä käyttäville tutkijoille. Haasteen voi muodostaa työmäärä, koska aineistot voivat olla erittäin laajoja.

Monimutkaisempi tapa pohjautuu erilaisiin käyttäjäkokeisiin, joilla pyritään analysoimaan ihmisten ymmärrystä aiheista ja vertaamaan niitä aihemallinnuksen kautta syntyneisiin aiheisiin (Chang ym. 2009; Towne ym. 2016). Esimerkiksi Chang ym. (2009) käytti viittä yleisintä sanaa luodakseen mielikuvan siitä, mistä aiheessa on kyse. Tämän jälkeen esitettäisiin useampaa vaihtoehtoa kuudenneksi sanaksi, minkä jälkeen voidaan suoraan arvioida, vastaako aihemallinnus ihmisen tulkintaa. Vaihtoehtoisesti voidaan näyttää kolme eri dokumenttia: kaksi samasta aiheesta ja kolmas muista aiheista (Towne ym. 2016). Jälleen on helppo arvioida, vastaavatko aiheet ihmisten mielekästä tulkintaa. Lähestymistapa on nopeahko suorittaa ja toimii hyvänä indikaattorina aiheiden luotettavuudesta. Valitettavasti menetelmä ei ole vakiintunut ja vaatisi tarkempaa perustutkimusta luotettavuuden osalta.

Viimeinen mahdollinen tapa on laadullisesta tutkimuksesta tuttu useamman tulkitsijan käyttö sekä heidän välinen keskustelu mahdollisista tulkinnoista. Tällöin useampi tutkija tarkastelee sanalistoja sekä muodostaa niille tulkintansa. Tämän jälkeen tulkinnoista keskustellaan sekä niitä täsmennetään kattamaan eri tutkijoiden näkökulmia. Lähestymistapa on jälleen yksinkertainen sekä laadullisten menetelmien käyttäjille tuttu. Toisaalta siinä oleva mitattavuuden puute voi aiheuttaa kritiikkiä epämääräisyydestä laskennallisten tieteilijöiden keskuudessa.

Mitä aihemalli tuottaa?

Tulkinnan lisäksi on syytä mietitä, mitä aihemallinnuksen löytämän aiheet ovat. Teknisestihän tuloksena on jokaiselle dokumentille jakauma eri aiheista sekä sanojen jakautuminen eri aiheisiin. Tällä tavoin voidaan määrittää esimerkiksi se, mihin aiheisiin kukin dokumentti kuuluu eniten. Kyseessä on kuitenkin varsin abstrakti määritelmä – aihemallinnus tuottaa yhteen liittyviä sanaryhmittymiä sekä dokumenttien ja aiheiden suhteita. Tämän takia aiheille onkin haettu monia teoreettisia vastinpareja: niiden on ajateltu olevan kehyksiä (*frames*), asioita (*issues*), teemoja (*theme*), diskursseja (*discourses*) tai draaman näytöksiä (*dramatistic scene*) (Jacobi ym. 2016; Mohr ja Bogdanov 2013).

Selvää on, että keskustelu aihemallin soveltuvuudesta ja yhdistettävyydestä yhteiskuntatieteelliseen teoriaan tulee jatkumaan (esim. Ylä-Anttila ym. 2018). Aihemallinnus laskennallisena suorituksena on samanlainen riippumatta siitä, mitä tulkintoja ja teoreettisia käsityksiä aiheille annetaan. Sen sijaan teoreettisen merkityksen anto aiheille on uskoakseni parhaiten sidottavissa tutkimuksen taustakirjallisuuteen ja sen käsitteellistyksiin. Esimerkiksi uutismedian tutkimuksessa kehykset ovat niin vakiintunut käsite, että aihemallinnuksen tulokset pyritään tulkitsemaan kehyksinä. Selvää on, ettei aihemallinnus tuota mitään käsitettä valmiina – tähän ohjatut menetelmät ovat usein parempia, koska niissä opetetaan luokitellun aineiston avulla teoreettisesti mielekäs tulkinta. Aiheiden analyysissä täytyy kuitenkin varoa liiallista argumentaatiota teoreettisten käsitteiden kautta.

OSATUTKIMUKSISSA KÄYTETTÄVÄ AINEISTO

Yksinkertaisuuden vuoksi molemmissa osatutkimuksessa käytetään samaa empiiristä aineistoa: suomalaisten puolueiden yleisohjelmia 1880-luvulta tähän päivään. Puolueohjelmat kerättiin Poliittisten ohjelmien tietovarannosta[5]. Yhteensä aineistoon kuului 198 puolueohjelmaa, joiden jakauma ajallisesti sekä puolueittain on esitetty tarkemmin taulukossa 2. Aineisto esikäsiteltiin lemmaamalla sekä poistamalla yleiset sekä harvinaiset sanat.

Puolueen nimi	Määrä
KD	16
KESK	24
KOK	23
KOM	5
RKP	10
SDP	10
SKDL	5
SKP	10
VAS	9
VIHR	11
Muut	71

(a) Aineisto puolueittain. Kaikki puolueet joilla aineistossa vähemmän kuin viisi ohjelmaa sijoitettu muut-kategoriaan.

Vuodet	Määrä
1880–89	1
1890–99	2
1900–09	8
1910–19	7
1920–29	8
1930–39	11
1940–49	9
1950–59	8
1960–69	17
1970–79	18
1980–89	21
1990–99	37
2000–09	37
2010	5

(b) Aineisto vuosittain.

Taulukko 2: Aineistona käytetyt puolueohjelmat

OSATUTKIMUS 1: AIHEMÄÄRÄN VALINNAN HAASTEET

Aihemallinnus on eräs laskennallisen analyysin keino tekstianalyysin suorittamiseen. Menetelmistä puhutaan usein pätevyyden ja uskottavuuden kautta. Pätevyys (validiteetti) liittyy tutkimuksen tekemiseen, eritoten siihen, että ilmiötä on kuvattu järkevästi. Uskottavuus (reliabiliteetti) taas liittyy siihen, että tutkimus on tehty johdonmukaisesti.

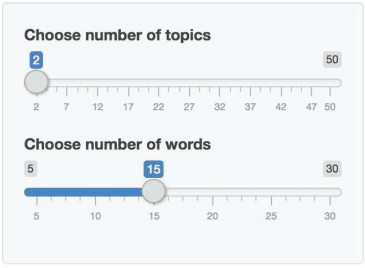
Määrällisessä tutkimuksessa pätevyyden keskeinen paino on onnistuneessa operationalisoinnissa ja otannassa: pätevässä tutkimuksessa on tarpeen mitata oikeaa asiaa sekä onnistua väitteiden yleistettävyydessä. Vastaavasti uskottavuus viittaa mittareiden kykyyn tuottaa samanlaisia tuloksia riippumatta esimerkiksi mittaajasta (Metsämuuronen 2003). Laadullisessa tutkimuksessa käsitteet pätevyys ja uskottavuus ovat herättäneet keskustelua ja muita käsitteitä on ehdotettu. Myös laadullisessa tutkimuksessa on pyrkimyksenä saavuttaa mielekäs kuvaus tutkimuksen kohteesta. Tutkimuksen tulisi vakuuttaa lukija analyysin ja johtopäätösten mielekkyydestä, esimerkiksi tuomalla esille tutkimusaineistoa tai kuvaamalla tulkitsijan lähestymistapoja ilmiöön. Pätevyydestä ja uskottavuudesta laadullisten ja määrällisten menetelmien osalta on kirjoitettu erittäin laajasti, eikä yllä oleva lyhyt kuvaus tee oikeutta tälle varsin laajalle keskustelulle.

Kuinka pätevyyttä ja uskottavuutta tulisi käsitellä ohjaamattomissa koneoppimismenetelmissä, kuten aihemallinnuksessa? Erityisesti on kysyttävä, kuinka tulisi huomioida se, miten tutkijoiden erilaiset lähestymistavat sekä kokemukset vaikuttavat harkintaan tulkittavasta aihemäärästä?

Menetelmä

Pyysimme neljää käyttäjäkokeeseen osallistujaa[6] valitsemaan aihemäärän, jolla aineistosta muodostuu selkein puoluekentää ja sen muutosta 1900-luvun alusta kuvaava kokonaisuus. Aihemallinnus olisi luonteva yhteiskuntatieteellinen menetelmä osallistujille esitetyn tutkimusongelman ratkaisemiseen. Kysymyksenasettelu on laadittu mahdollisimman samankaltaiseksi yhteiskuntatieteellisessä tutkimuksessa käytettyjen aihemallinnuksen sovellusten kanssa, vaikkakin aihemallinnuksen aiheille ei ole pyritty antamaan täsmällisempää teoriasta ja käsitteistä juontuvaa merkitystä. Kuten yllä kuvataan, tämä vaihe seuraa usein aihemallinnusta valittujen aiheiden analyysin jälkeen.

Osallistujilla oli käytössään vuorovaikutteinen visualisaatio-ohjelma[7]. Kuten kuva 2 näyttää, osallistujat pystyivät vuorovaikutteisesti vaihtamaan aihemäärää sekä tutkimaan kunkin aihemäärän 15 yleisintä sanaa. Sanalistat valittiin aiheiden esittelymuodoksi, koska niiden on havaittu tuottavan selkeimpiä ja johdonmukaisimpia tulkintoja verrattuna esimerkiksi pelkkien tekstiesimerkkien käsittelyyn (Aletras ym. 2017). Vaikka aihemallinnus suoritettiin laajalle aihemäärälle (k=2–300), käyttäjäkokeesen valittiin tarkasteltavaksi vain 20 eri aihemäärää (k=20–39). Pienemmällä aihemäärällä haluttiin rajata osallistumiseen kuluva aika sekä vähentää käyttäjäkokeen kognitiivista kuormaa osallistujille. Aihemäärän väli valittiin kirjallisuudessa suositellun harmonisen loglikelihood-mittarin mukaan siten, että sen suosittelema aihemäärä kuului tarkasteluvälille.



Words per topic

	Topic 1	Topic 2
1	ihminen	maa
2	suomi	valtio
3	yhteiskunta	työ
4	kaikki	toiminta
5	ympäristö	yhteiskunta
6	työ	kansa
7	palvelu	taloudellinen
8	tavoite	kaikki
9	mahdollisuus	turvata
10	kehitys	kehittää
11	ongelma	saada
12	oikeus	mahdollisuus
13	uusi	kunta
14	arvo	puolue
15	kansalainen	tuotanto

Kuva 2: Visualisointiohjelma. Vasemmalla olevasta valikosta voidaan valita aihemäärä. Ohjelma näyttää tälle aihemäärälle ja jokaiselle aiheelle 15 kuvaavinta sanaa.

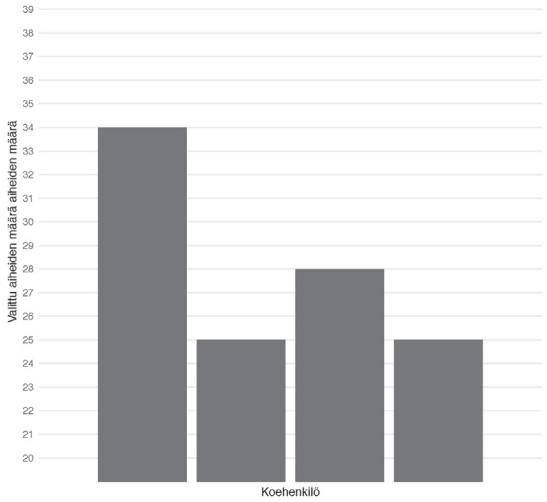
Tarkistaaksemme osallistujien ymmärryksen aihemallinnuksesta pyysimme näitä tutustumaan menetelmään yleistajuisen englanninkielisen artikkelin avulla (Blei 2012). Artikkelin ja menetelmän sisäistäminen tarkastettiin kolmella yleistajuisella väitelauseella. Kaikki käyttäjäkokeeseen osallistujat osoittivat hallitsevansa aihemallinnuksen perusteen. Lisäksi osallistujat olivat yhteiskuntatieteen alalta vähintään ylemmän korkeakoulututkinnon suorittaneita sekä olleet tutkijakoulutuksessa. Käyttäjäkokeeseen osallistujat siis olisivat voineet tehdä vastaavilla menetelmillä ja aineistoilla analyysiä oman tutkimuksensa osalta. On tietenkin ilmeistä, että kukin koehenkilö lähestyy toisaalta aineiston tulkintaa ja toisaalta avoimeksi jätettyä tehtävänantoa omien kokemuksiensa ja näkökulmansa kautta[8].

Tämän takia pyysimme osallistujia sekä valitsemaan parhaimman aihemäärän että perustelemaan, kuinka he päätyivät ehdottamaansa aihemäärään. Näitä vastauksia analysoidaan kevyen laadullisesti; koska osallistujia on vain neljä, ei ole mielekästä pyrkiä tarkkaan analyysiin. Perusteluiden kautta voidaan kuitenkin kuvata myös sitä, kuinka tietty aihemäärä on koettu kysymykseen sopivaksi ja mitä strategioita osallistujat käyttivät päätöksenteon tukena.

Sisällöllistä tulkintaa korostavan menetelmän lisäksi voidaan käyttää myös tilastollisia menetelmiä aihemäärän valintaan, kuten yllä on esitetty. Käytämme näitä menetelmiä ja arvioimme tällä perusteella parhaimman aihemäärän sekä vertaamme sitä tutkijoiden ehdottamiin sisällöllisiin aihemääriin.

Tulokset

Osallistujien mielestä sopivin aiheiden määrä oli välillä 25 ja 34, kuten kuvasta 3 näkyy. Aihemäärien jakauma ei ole painottunut, vaan enemmänkin satunnainen. Yllättäen kaksi neljästä osallistujasta päätyivät suosittamaan samaa aihemäärää, 25 aihetta. Kaikin puolin yksinkertainen käyttäjäkoe nostaa esille jo haasteita tulkintaa korostavissa aihemäärän valinnoissa: kuinka tämä subjektiivinen valinta tulisi perustella ja mitä merkitystä aihemäärän valinnalla on tutkimuksen reliabiliteetille?



Kuva 3: Koehenkilöiden mielestä sopivimmat aihemäärät

Ymmärtääksemme tarkemmin, mikä voisi selittää eroja aihemäärissä, kysyimme osallistujilta heidän kriteereistä ja strategiasta aiheen valintaan. Tunnistimme, että osallistujat käyttivät kahta eri strategiaa sopivan aihemäärän valitseminen; aihepiiristä olemassa olevaa tietoa ja ennakkokäsityksiä korostavaa tai aihemallinnuksen selkeyttä korostavaa strategiaa. Tarkastelemme myös erikseen samaan aihemäärään päätyneiden osallistujien perusteluita.

Ensimmäisessä strategiassa käytettiin olemassa olevia ennakkokäsityksiä aiheiden järkevyyden ja mielekkyyden tarkasteluun. Tätä strategiaa käyttämällä aiheille annettiin poliittisesti mielenkiintoiset tulkinnat ja niitä käytettiin apuna arvioidessa aihemalleja. Esimerkiksi osallistuja kuvasi, kuinka ”Maalaisliitolla oli mielestäni esimerkiksi enemmän maatalouteen ja sitten aluepolitiikkaan ja pienyrittäjyyteen liittyvät kaksi topiikkia. Maalaisliitto toimi nyt tässä ikään kuin proxyna.”

Tätä käyttäen hän valitsi aihemäärän, joka toi esille niin Maalaisliiton eri poliittiset painoalueet kuin myös muita puolueita. Osallistuja päätyi 34 aiheeseen. Samoin saatettiin korostaa tiettyjen puolueiden tai politiikka-aiheiden puuttumista analyysistä ja arvioida aihemalleja, sillä perusteella, mitä ne paljastavat ilmiöstä: ”pienemmällä aihemäärällä joukosta vaikuttaa puuttuvan joitain aiheita, jotka vaikuttavat relevanteilta, esim. Piraattipuolue-aihe.”

Toinen strategia taas korosti aiheiden vertailua keskenään ja ylimääräisten tai päällekkäisten aiheiden välttämistä: ”suuremmalla määrällä taas mukana alkaa olla epärelevantin näköisiä aiheita sekä keskenään samaa asiaa koskevia aiheita”; ”tuntui että 39 erotteli jossain määrin turhankin tarkasti tuottamalla esim. monta omaa topicia tietyille puolueille”; ”noin 25 tienoille asti tuntuu, että topiikit pysyvät selkeämmin erillisinä, sen jälkeen alkaa tulla päällekkäisyyttä ja on vaikeampi tulkita mitkä [ovat] topiikkien eroja.”

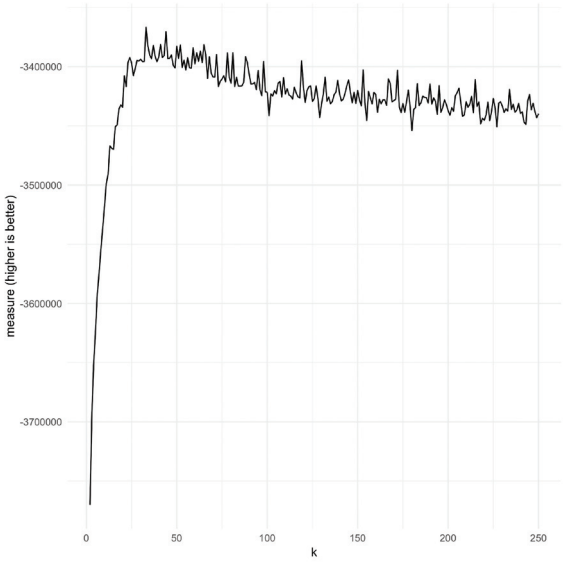
Kaksi osallistujaa mainitsivat käyneensä läpi aiheita systemaattisesti hakemalla aihemäärän kautta väliä, jossa aihemallinnuksen tulokset olivat selkeimpiä. Osallistuja esimerkiksi kuvasi tarkastaneensa ensin molemmat ääripäät (20, 39) ja puolivälin (30). Hän kommentoi, että ”30 aiheen mallinnus puolestaan oli jo melko selkeä verrattuna 20 aiheen mallinnukseen, jossa oli melko monta puurotopicia.”

Tämän perusteella osallistuja päätti vielä tarkastaa näistä puolivälin (25) ja totesi, ettei siinä ”hänen mielestään hukattu mitään olennaista.”

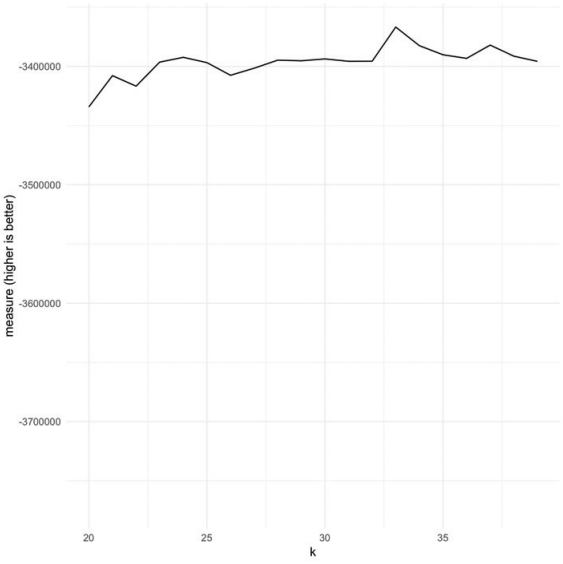
Samaan aihemäärään (25) päätyneet osallistujat perustelivat aihemäärää samalla tavoin: aiheet olivat tarpeeksi erillisiä ja erottelivat aineistoa tarpeeksi, eikä niissä ei ollut päällekkäisiä aiheita. Toisaalta samankaltaisella perustelulla päädyttiin myös 28 aiheeseen. Vaikka alustavasti tulokset voisi tulkita positiivisiksi – osa osallistujista päätyi subjektiivisella arvioinnilla samaan tulokseen – tulokset osoittavat myös eroja strategian käytössä. Tämä tuo esille, ettei saman valintastrategiankaan käyttö välttämättä takaa samoja tuloksia subjektiivisessa tulkinnassa. Aihemäärän valintaa subjektiivisesti ei siis voida pitää kovinkaan luotettavana lähestymistapana. Mahdollisesti aihemallinuksen tulkinnan avaaminen ja läpinäkyvyys – esimerkiksi ottaen mallia laadullisesta tutkimuksesta – voisi selkeyttää tätä tilannetta, mutta aihemallinnusta soveltavissa töissä näin tehdään valitettavan harvoin. Työn toinen osatutkimus osoittaa, että aihemäärän valinta vaikuttaa analyysin erottelukykyyn.

Tuskin on yllättävää, että käyttäjäkokeisiin osallistujien ehdottamat lukumäärät eroavat, eivät vain keskenään, vaan myös aineistoista laskettavista tilastollisista mittareista. Kuten yllä on kuvattu, myös tilastollisissa mittareissa on eroja, minkä lisäksi suositeltu tilastollinen mittari

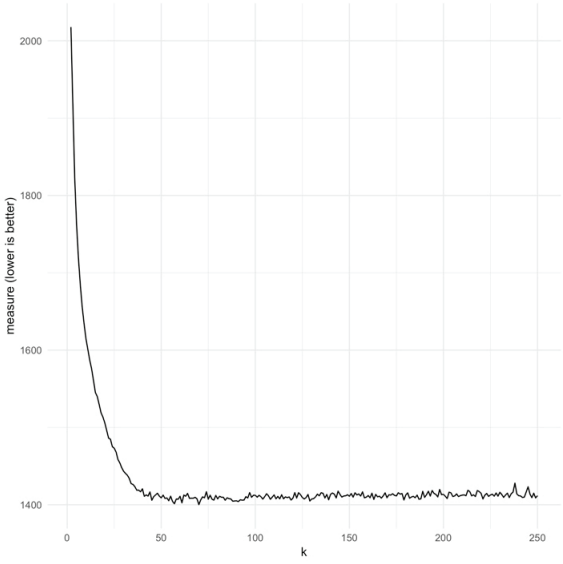
vaihtelee kirjallisuudessa. Paras aihemäärä on harmonisen loglikelihood-arvon perusteella 33 ja perpleksiteetti-arvon mukaan 70. Kuten kuvat 4a sekä 4b näyttävät, loglikelihood-arvo paranee merkittävästi 25 aiheeseen asti ja alkaa huononemaan 50 aiheen jälkeen. Toisaalta perplexiteetti (kuvat 4c sekä 4d) paranee noin 50 aiheeseen asti, jonka jälkeen sen muutokset tasaantuvat.



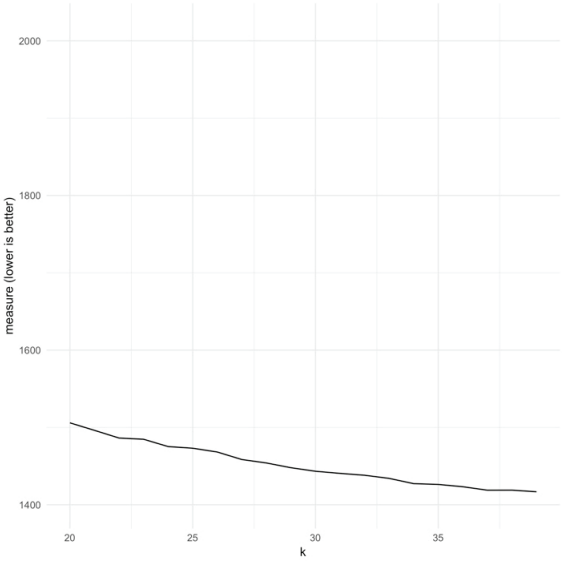
4a: Harmoninen loglikelihood – tarkastelu *k*n arvoilla 2–250



4b: Harmoninen loglikelihood – tarkastelu *k*n arvoilla 20–30



4c: Perpleksiteetti – tarkastelu *k*n arvoilla 2–250



4d: Perpleksiteetti – tarkastelu *k*n arvoilla 20–30

Kuva 4: Aiheiden määrän sopivuus eri laskennallisilla mittareilla.

Keskustelu

Pienimuotoinen käyttäjäkoe osoittaa, ettei yhteiskuntatieteissä käytetty, sisällöllistä tulkintaa korostava menetelmä aiheiden valintaan ole erityisen luotettava. Aihemäärät eroavat tutkijoiden välillä, mikä ohjaamattoman koneoppimisen tapauksessa johtaa erilaisiin tuloksiin (tätä käsitellään tarkemmin seuraavassa osatutkimuksessa). Myös strategiat aiheiden määrän valintaan ovat erilaisia. Toiset korostivat ennakkokäsitysten ja toiset erottelukyvyn merkittävyyttä aihemäärän valinnassa. Molemmat lähestymistavat ovat perusteltuja, eikä siis ole ilmeistä, onko mikään käyttäjäkokeeseen osallistuneiden ihmisten suosittelema lukumäärä 'paras' kuvaamaan aineistoa.

Toisaalta, myös erilaiset tilastolliset mittarit tuottivat erilaisia suosituksia aihemäärästä, kuten kuvassa 4 havaittiin. Tilastollisten mittarien osalta käydään myös aktiivista keskustelua tietojenkäsittelytieteessä eri aiheenvalinnan menetelmien laadusta. Uusin kirjallisuus on päätynyt suosittelemaan harmonista loglikelihood-arvoa, mutta sekään välttämättä tuottaa parhainta tulosta (esim. Griffiths ja Steyvers 2004; Wallach ym. 2009). Samalla tavoin kuin subjektiivisen tulkinnan malleissa voidaan siis kysyä, että onko näilläkään mittareilla mahdollista selvittää "parasta" aihemäärää kuvaamaan aineistoa.

Tulokset osoittavat, että eri lähestymistavat voivat johtaa hyvinkin erilaisiin tuloksiin. Tässä tapauksessa löytyi viisi eri vaihtoehtoa parhaaksi aihemallinnukseksi aineistolle. Jokainen niistä on perusteltavissa ja ne olisivat luultavasti hyväksyttyjä myös tieteellisissä julkaisuissa[9]. Tästä seuraa haastava dilemma: mikä vaihtoehdoista tulisi valita jatkoanalyyseihin parhaiten sopivana? Aihemallinnuksen lisäksi vastaava haaste on olemassa kaikissa ohjaamattoman koneoppimisen menetelmissä: niissä tutkijan tulee aina tehdä jotain rajauksia ja valintoja. Tämä on laskennallisen data-analyysin iso haaste. Kuten Watts (2011) argumentoi, ihmiset ovat erittäin hyviä muodostamaan mielekkäitä tulkintoja *kaikista* tuloksista. Tällöin erilaiset aihemäärät voivat vaikuttaa selkeiltä ja järkeviltä, vaikka aineiston "todellinen" aihejakauma olisi täysin toisenlaisen. Tutkimusmenetelmän toistettavuuden kannalta tämä on erittäin haastavaa, ja siksi subjektiivisten arvioiden sijaan olisikin syytä käyttää tilastollisia mittareita, jotka tuottavat toisinnettavan tuloksen. Toisaalta, tilastollisella mittarilla luodun aihemallinnuksen tulkinnallisuus voi olla haastavampaa kuin subjektiivinen arvion luomiin aiheisiin. Siksi ohjaamattomia menetelmiä soveltaville yhteiskuntatietelijöille suositellaan:

Suositus 1 Käytä laskennallisia mittareita mallin parametrien (kuten aiheiden määrän) valinnassa. Kuten jo yllä huomioitiin, tähän on tarjolla useita erilaisia mittareita (Wallach ym. 2009). Sopivan mittarin valinta vaatii siis alan kehityksen seuraamista. Tätä tekstiä kirjoittaessa monet ovat suositelleet harmooniseen keskiarvoon perustuvaa logaritmisen uskottavuuden -mittaria (harmonic loglikelihood).

Tehtyä pienimuotoista käyttäjäkoetta voidaan toki kritisoida monilla tavoin. Selkein kritiikki on pieni osallistujamäärä, minkä takia ei ole mahdollista tehdä laajempia tilastollisia yleistyksiä aihemäärästä. Toisaalta, koehenkilöillä ei ollut käytössä laajempaa tai yhtenäisempää teoriataustaa mihin tulos sidottaisiin: tutkimuskysymyksenä suomalaisen puoluekentän muutosten ymmärtäminen 1900-luvulta tähän päivään on erittäin avoin. Tällöin valitut aihemäärät voivat myös kuvastaa eroavaisuuksia siinä, minkälaiseen kirjallisuuteen ja taustaan koehenkilöt sijoittivat

tehtävänannon. Samoin muut taustatekijät ja esimerkiksi mielikuvat Suomen puolueohjelmien sisällöistä voivat hyvin ohjata tässä vaiheessa tapahtuvaa tulkintaa. Toisaalta, jos tutkimuskysymys olisi ollut erilainen, olisiko se johtanut erilaisiin strategioiden käyttöihin? Jos menetelmät ja tulokset eroavat merkittävästi, niin mitä tämä kertoo ohjaamattoman menetelmän mahdollisuuksista tuottaa toistettavia tuloksia? Eräs mahdollisuus olisikin käyttää ohjaamattomia menetelmiä vain aineiston ymmärtämiseen, mutta käyttää ohjatun koneoppimisen menetelmiä yhdistäessä aineistoa yhteiskuntatieteellisiin käsitteisiin, eli opettaa tietokonetta tunnistamaan esimerkkien pohjalta nämä käsitteet (vrt. Nelson 2017).

OSATUTKIMUS 2: AIHEMÄÄRÄN VAIKUTUS EMPIIRISIIN LÖYDÖKSIIN

Osatutkimuksessa 1 osoitettiin, että aihemäärän valinta on kaikkea muuta kuin jo ratkaistu ongelma. Osatutkimus 2 tarkastelee eri aihemäärillä luotujen aihemallien eroja mahdollisessa jatkoanalyysissä. Osatutkimuksen 1 tutkimuskysymystä mukaillen tässä osatutkimuksessa tavoitteena on kuvailla suomalaisen puoluejärjestelmän muutosta 1900-luvulta tähän päivään. Aiheesta on luontaisesti kirjoitettu erittäin runsaasti, ja seuraava lyhyt katsaus kirjallisuuteen ei tee kunniaa olemassa olevalle tutkimukselle. Tämän lyhyen katsauksen avulla voidaan kuitenkin erottaa muutama selkeä havainto ja oletus puoluekentän muutoksesta: niiden avulla voidaan tarkastella eri aihemäärien vaikusta tutkimuksen tuloksiin.

Aikaisempi tutkimus puoluejärjestelmien kehittymisestä

Puoluekentän muutosta on pyritty selittämään puolueiden toimintalogiikan muutosten kautta. Farrell ja Webb (2000) kuvaavat kuinka poliittiset puolueet ovat muuttuneet hajautuneista liikkeistä keskitetysti hallittuihin organisaatioihin. Erityisesti laaja muutos on tapahtunut kollektiivi-identiteettien sijaan mahdollisimman laajoiksi yleispuolueiksi (*catch all* -puolueiksi; esim. Scarrow 2000). Sellaisenaan yleispuolueen määritelmää on pidetty haastavana (esim. Maas 2001; Krouwel 2003). Kuitenkin eräs keskeinen piirre määritelmässä on ideologisen pohjan laajentuminen sekä keskittyminen useampiin politiikan alueisiin (esim. Maas 2001; Krouwel 2003). Yleisesti on argumentoitu, että yllä kuvatut muutokset intressipuolueista yleispuolueiksi ovat ilmeisiä kaikissa länsimaalaisissa demokrati

Yllä oleva kirjallisuus painottui kansainväliseen tilanteeseen. Samankaltaisia tuloksia on tuotu esille myös keskittyen vain suomalaisiin puolueisiin. Esimerkiksi hiljattain julkaistu Mickelsson (2015) jäsentää puoluejärjestelmän kehitystä kuudessa ajanjaksossa. Ensimmäinen ajanjakso (1905–1922) keskittyi kansakunnan sekä valtion järjestäytymiseen. Poliittiset aiheet keskittyivät muun muassa kielikysymyksen, työväen sekä maaseudun aseman ympärille. Toinen ajanjakso (1923–1939) korosti työväen ja porvareiden välistä eroa sisällissodan seurauksena. Intressiryhmät nousivat edelleen esille kolmantena ajanjaksona (1940–1965), jota Mickelsson (ibid.) kutsuu myös taistelevien intressipuolueiden Suomeksi. Puolueet ajoivat tarkemmin omien intressiryhmiensä etuja, tosin ne kehittyivät kohti yleispuolueita. Intressiryhmien sijaan ideologiat korostuivat neljännellä jaksolla (1966–1978), jota Mickelsson (ibid.) kuvaa myös rajujen

muutosten ajaksi puoluekentällä. Muutokset kumpusivat nuorisoryhmien kautta, esimerkiksi ihmisoikeustyön ja pasifismin kautta. Viides ajanjakso (1979–2007) korosti puolueiden toiminnan muutokseen media- ja markkinapuolueiksi. Modernisaatio on nähtävissä esimerkiksi ekologisen puoluekentän kehittymisenä. Politiikan aiheissa oli murros 'uuteen politiikkaan', elämäntapoihin, feminismiin sekä globaaliin tasa-arvoon. Murros jatkui kuudennella ajanjaksolla (2008–2015) kun vihreiden uuden politiikan näkökannoille muodostui poliittinen vastavoima perussuomalaisista. Arter (1999) on vastaavasti analysoinut keskustan historiallista muuttumista yleispuolueeksi. Hän näki keskeisenä muutoksena keskustan pyrkimyksen laajentaa ideologista pohjaansa vuoden 1962 puolueohjelmassa. Lisäksi puolueen kannattajakunta monipuolistui neljännen ajanjakson aikana (1966–1978). Näin ollen Mickelssonin (2015) esittämät ajatukset puolueiden muutoksesta kyseisellä aikakaudella saavat tukea myös muusta tutkimuksesta ja onkin ilmeistä, että myös suomalaisessa puoluekentässä on yleispuolueita.

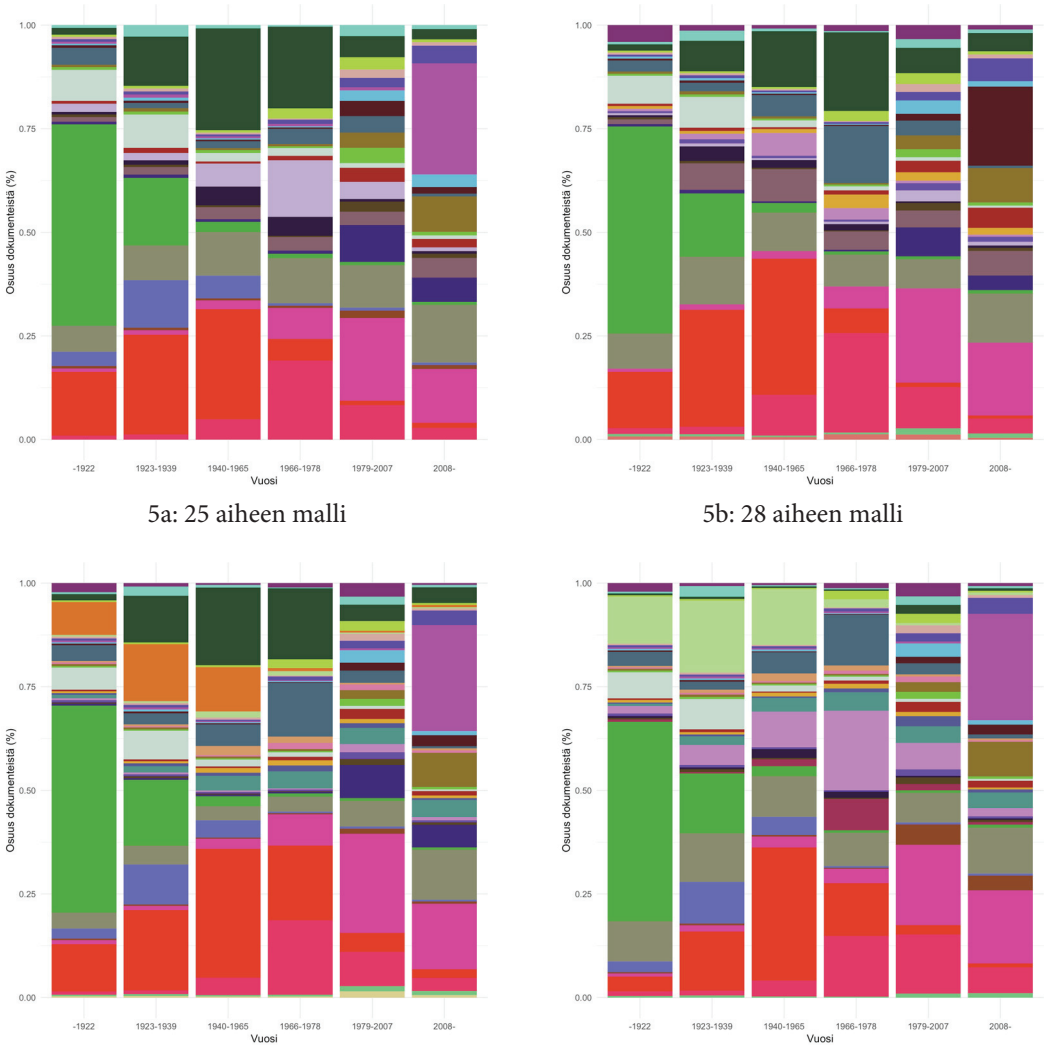
Yllä olevan kirjallisuuden perusteella on ilmeistä, että myös Suomen puoluekentässä on havaittu siirtyminen intressipuolueista yleispuolueiksi 1900-luvun aikana, erityisesti vuosina 1966–1978. Käytämme tätä historiallisen kehityksen muutosta arvioimaan aihemallien luomia tuloksia ja niiden tarkkuutta. Tarkastelemme tätä kysymystä seuraavaksi tarkemmin puolueohjelmien aihemallinnuksen valossa.

Analyysi ja tulokset

Ensimmäinen vaihe aihemalleihin perustuvissa analyyseissä on antaa aiheille merkitys, eli tulkita aihemallien aiheet mielekkäästi. Tulkinnassa voidaan havainnoida aiheissa yleisiä sanoja tai sanalistoja. Lisäksi aiheista voidaan valita edustavia dokumentteja. Vain sanalistojen tulkinnan on osoitettu tuottavan erittäin hyviä aihetulkintoja (Aletras ym. 2017). Tällöin edustavien dokumenttien valintaa voidaan käyttää validointivaiheessa. Liitetaulukossa 1 esitetään tulkinnat eri aiheiden määrillä tiivistetyssä muodossa. Tulkinnat perustuvat vain sanalistojen käyttöön, eli aihemallin tuloksia ei ole tätä tarkemmin validoitu.

Aihemallinnuksessa voidaan edelleen jalostaa mielekkäämpiä kokonaisuuksia. Yhteiskuntatieteissä tämä tapahtuu usein nojaamalla teoreettisiin käsitteisiin. Esimerkiksi aiheita voitaisiin ryhmitellä edelleen kokonaisuuksiksi (vrt. Strauss ja Corbin 1990). Tässä työssä tämä ei ole tutkimuskysymyksen kannalta mielekästä: aihemallien ryhmittymät kuvaavat jo puolueita sekä politiikka-aiheita. Toisaalta, vaikka samalle puolueelle aihemallinnuksen kautta voidaan tunnistaa useampia aiheita (esimerkiksi kokoomus liitetaulukossa 1), ovat nämä sisällöllisesti varsin erilaisia eikä siinä mielessä niiden yhdistäminen ole mielekästä; tutkimuskysymyksenä kun on tarkastella puolueohjelmien muutosta sekä aihemäärien vaikutusta tähän ongelmaan. Toisenlaisella tutkimuskysymyksellä kuitenkin yhdistäminen voi olla tarpeellista ja mielekästä.

Ensimmäinen viite aihemäärän valinnan merkityksestä on nähtävissä liitetaulukosta 1: eri aihemäärät johtavat eri tulkintoihin puolueohjelmien sisällöstä ja niiden aiheista. Jotkin aiheet ovat vakaita ja esiintyvät kaikissa aihemäärissä. Toiset aiheet taas tulevat esille, kun aihemäärää kasvatetaan ja mallin "erottelukyky" kasvaa.



5a: 25 aiheen malli

5b: 28 aiheen malli

5c: 34 aiheen malli

5d: 33 aiheen malli

Kuva 5: Aiheiden jakautuminen puolueohjelmissa. Aiheet väritetty tulkintojen mukaan yhteneväisesti kuvaiden välillä. Tulkinnat esitetty taulukossa L1.

Aihemallinnus auttaa luomaan kokonaiskuvan Suomen puoluekentän muutoksista: kuva 5 näyttää tunnistettujen aiheiden esiintymisen puolueohjelmissa tarkasteltavien aihemäärien muuttuessa käyttäen Mickelssonin (2015) esittämää kuuden aikakauden mallia. Kuten kuvista näkyy, pääsääntöisesti aihemallien ryhmittelyt ja antamamme tulkinnat tuottavat samankaltaisen kuvauksen Suomen poliittisesta historiasta. Samaan aikaan eri mallien kautta muodostuva kuva puolueiden historiasta sisältää ilmeisiä eroja. Esimerkiksi ensimmäiselle, vuoteen 1922 yltävälle ajanjaksolle 25 ja 28 aiheen mallit (kuvat 5a ja 5b) eivät tuo esille maalaisliiton muutosta keskustaksi, kun taas 33 ja 34 aiheen malleissa (kuvat 5d sekä kuva 5c) aihe näkyy selvemmin (kuvion värit vihreä ja vaalean vihreä).

Poliittisten puolueiden ja kentän murrokset (ajanjaksot: 1979–2007 sekä 2008–) ovat selkeämmin esillä myös 33 ja 34 aiheen malleissa. Kuten kuva 5d näyttää, murrosajanjaksolle

on ilmeistä useiden eri poliittisten aiheiden esiintyminen, jotka ovat olleet marginaalissa aikaisempina vuosina. Tämä muutos on vähemmän ilmeinen 25 aiheen mallissa (kuva 5a), jossa väriskaalan muutos on vähäisempi. Vastaavia pienehköjä eroja aiheiden esiintymisessä on nähtävissä enemmän. Tämä antaa lisää osviittaa aiheiden määrän merkityksestä aihemallinnusprosessissa. Kuten näytimme, eri aihemäärät voivat johtaa erilaiseen tulkintaan taustalla olevasta ilmiöstä, tässä tapauksessa poliittisesta historiasta. Palaamme tähän tarkemmin osatutkimuksen keskusteluosiossa.

Mutta onko puolueohjelmien tarkastelun avulla mahdollista nähdä muutos intressipuolueista yleispuolueiksi? Ensimmäinen vaihe kysymykseen vastaamisessa on luonnollisesti *operationalisoida* käsitteinä toisaalta intressipuolue ja toisaalta yleispuolue. Yleispuolueelle ja intressipuolueelle on tarjottu useita erilaisia kriteereitä: yleispuoluetta kuvaa niin niiden sisäinen organisoituminen kuin suhde äänestäjäkuntaan (Kirchheimer 1990; Krouwel 2003; Maas 2001). Puolueohjelmien analyysissä eräs keskeinen määritelmä on, että intressipuolueilla on selvä ja pieni joukko aiheita, joissa puolue on aktiivinen, kun taas yleispuolueiden puolueohjelma käsittelee useampia aiheita ja ottaa laajemmin kantaa yhteiskuntaan. Tämä on mahdollista edelleen operationalisoida aihemallinnuksen kautta: yleispuolueen puolueohjelma ottaa enemmän ja selkeämmin kantaa useaan aiheeseen, kun taas intressipuolue keskittyy tarkemmin omalle kannattajakunnalle keskeisiin teemoihin.

Kirjallisuuden mukaista kehitystä on tapahtunut: aiheiden määrän kasvu näkyy kuvassa 5 värien määrän kasvuna kaikissa malleissa ensimmäisen ja vuoteen 1978 päättyvän neljännen ajanjakson välillä. Visuaalisen tarkastelun lisäksi ilmiötä voidaan tarkastella aihejakauman muutoksien kautta. Vaikka yksittäinen aihe ei *aina* kuvaa tiettyä intressiryhmän etujen ajamista – keskeinen piirre yleispuolueen ja intressipuolueen välillä – liitetaulukko 1 osoittaa, että monet aiheet ovat nähtävissä myös tätä kautta. Poikkeuksen tähän muodostaa kunkin puolueen erityiset puolueaiheet, joissa jokainen puolue on varsin hyvin edustettuna. Koska aiheilla on myös intressiryhmiin liitettäviä merkityksiä, aihemallinnuksen aiheita voi käyttää arvioimaan puolueiden halua viestittää olevansa intressipuolueita.

Laskemme siis kunkin puolueen osalta niitä koskevien aiheiden merkittävyyden. Merkittävyyden kriteerinä olemme käyttäneet 10% rajaa: jos puolueohjelma mainitsee tietyn aiheen yli 10% arvosta, pidetään aihetta merkittävänä puolueohjelmalle[10]. Tämän kaltainen rajaus on tarpeen, koska aihemallinnuksessa jokainen aihe on jollain osuudella osana jokaista puolueohjelmaa. Tämän jälkeen vertaamme kaikkia puolueohjelmia keskenään sen perusteella, kuinka monta teemaa kullakin puolueella on esillä. Haasteen analyysiin tuo toki puolueiden sisäisen kehittyminen: onko mielekästä verrata 1980-luvun liikkeitä samalla tavalla kuin jo vakiintuneita toimijoita? Mickelssonin (2015) mukaan kehitys intressipuolueista laajemmiksi ideologiapuolueiksi tapahtui vuosina 1966–1978. Lisäksi, 1980-luvulla on samaan aikaan ollut olemassa sekä laajoja yleispuolueita että yhden asian liikkeitä, kuten ympäristöryhmittymiä, mikä voisi haastaa analyysiä[11]. Tämän takia tässä aineistossa rajaudutaan toisaalta arvioimaan muutosta ennen vuotta 1966 sekä toisaalta vuosien 1966 ja 1978 välissä merkittävien aiheiden määrän kannalta.

Taulukosta 3 nähdään, että puolueohjelmien merkittävien aiheiden määrä on kasvanut jälkimmäisellä tarkastelujaksolla. Tämä tukee ajatusta siirtymisestä kohti yleispuolueita, ei rajattuihin teemoihin keskittyviä intressipuolueita. Taulukko myös osoittaa aihemallinnuksen aihemäärän merkityksen tuloksen kannalta. 25, 33 ja 34 aiheen malleilla ero on tilastollisesti

merkittävä Wilcoxonin mittarilla[12] kun taas 28 aiheen mallilla tilastollista eroa ei ole. Tilastollisesti merkitsevät mallit myös esittelevät ilmiötä osittain eri tavoin: 25 aiheen mallilla ero on huima: vuosina 1966–1978 puolueet käsittelivät yhtä aihetta enemmän kuin aikaisemmin, kun taas 33 ja 34 aiheen malleilla ero on noin puolen aiheen verran. Taulukon 3 osittain ristiriitaiset tulokset tuovat erinomaisesti esille aihemallinnuksen haasteen käytännön tutkimuksessa. Aihemallinnuksen aikana tehdyt valinnat tuovat hyvinkin erilaisia näkökulmia Suomen puoluekentän muutokseen ja pahimmillaan johtivat (28 aiheen tapauksessa) tulokseen, ettei muutosta intressipuolueesta yleispuolueeksi ole havaittavissa tällä tavoin operationalisoituna. Kuitenkin, toiset aihemallinnuksen tulokset tunnistavat puolueiden kehittymisen: puolueohjelmat ottavat merkittävästi kantaa useampiin aiheisiin vuosina 1966–1978 kuin aikaisempina vuosina, eli puolueet ovat muuttuneet yleispuoluemaisiksi toimijoiksi.

Malli	**–1965**	**1966–1978**		
25 aihetta	2.41	3.34	$p = < 0.01$	($W = 413$)
28 aihetta	2.44	2.86	$p = 0.12$	($W = 622.5$)
33 aihetta	2.64	3.41	$p = 0.006$	($W = 516$)
34 aihetta	2.71	3.28	$p = 0.03$	($W = 567.6$)

Taulukko 3: Puolueohjelmassa olevien merkittävien aiheiden määrän keskiarvo ennen vuotta 1965 ja 1966–1978.

Keskustelu

Tärkein osatutkimuksen havainto on, että aiheiden jakaumien kautta esille tuodut muutokset ovat mahdollista sijoittaa Mickelsson (2015) esittämään jakoon Suomen puolueiden kehityksestä. Aihemallinnuksen kautta on selkeästi nähtävissä esimerkiksi muutos maaseutuyhteiskunnasta teolliseen yhteiskuntaan. Samoin myös muutos intressipuolueesta yleispuolueeksi on mahdollista operationalisoida ja tätä kautta mitata aihemallinnuksen kautta.

Koska yllä olevat muutokset olivat laajasti tiedossa kirjallisuudessa, luvun varsinainen anti onkin menetelmällinen. Jatkaen ensimmäisessä osatutkimuksessa 1 tehtyä kysymyksenasettelua, miten aihemäärä vaikuttaa analyysin tuloksiin? Niin kuva 5 kuin taulukko 3 näyttävät, että aihemäärän valinta on keskeisessä osassa tutkimusprosessia ja vaikuttaa tuloksiin. Nykyisillä tavoilla tehdä aihemallinnusta voitaisiin puolustaa jokaista tässä analyysivaiheessa käytettyä aihemäärää: 25, 28, 34 tai 33 aihetta. Tämä johtaa varsin haastavaan kysymykseen: mikä kuvista 5a, 5b, 5c tai 5d on ’oikea’ ilmentymä puolueohjelmista ja niiden historiallisesta kehityksestä. Välttääksemme tätä ongelmaa, esittelimme jo edellä suosituksen 1 tilastollisten mittarien hyödyntämisestä aihemäärän valinnassa. Näin esimerkiksi vältetään valintatilanne samanlaisella strategialla muodostettujen 25 ja 28 aiheen välillä. Molemmat valinnat olivat yhtä hyvin perusteltuja, mutta vain toinen niistä toi esille tilastollisesti merkittävän eron puolueohjelmien aiheiden määrän (ks. taulukko 3). Tätä kautta kysymys muutoksesta intressipuolueesta yleispuolueeksi saa erilaisia vastauksia.

Tulkinnallisuutta korostavaa lähestymistapaa puolustavat voisivat argumentoida, että eri aihemäärien avulla laskettujen aihemallien tuottamat erot ovat vähäisiä ja siksi tulkittavuuden korostaminen on hyväksyttyä. Tämä argumentti ei ole ilmiselvästi virheellinen. Esimerkiksi kuva 5 tuottaa samanlaisia trendejä ilmiöistä ja laajat, yhteiskuntatason merkittävät muutokset – kuten puolueiden aihemäärän laajentuminen vuoteen 1979 asti sekä uusien poliittisten liikkeiden nousu vuosina 1979–2007 – ovat esillä kaikissa kuvissa. Kuitenkin taulukko 3 tuo tarkemmin esille lähestymistavan ongelman: löydösten tilastollinen merkitsevyys ja vahvuus voivat muuttua aihemäärän valinnan seurauksena.

Löydös sellaisenaan jatkaa käynnissä olevaa varsin vilkasta keskustelua aihemallinnuksen eri vaiheiden, kuten esiprosessoinnin (Denny ja Spirling 2018; Schofield ja Mimno 2016) sekä mallinnuksen parametrien valinnan vaikutuksesta (Wallach ym. 2009) aihemallinnuksen lopputuloksiin. Tässä työssä esitellyt löydökset nostavat esille jälleen yhden uuden haasteen aihemallinnuksen soveltamiseen empiirisessä tutkimuksessa. Osatutkimus 2 osoittaa, että aihemäärän valinnalla on merkitystä tutkimuksen kannalta.

Toisaalta osatutkimus 2 tuo myös esille tulkinnan ja luokittelun haasteet aihemallinnuksessa. Esimerkiksi tässä työssä on päätetty esittää kukin aihe erillisenä kokonaisuutena. Kuten jo yllä keskusteltiin, olisi myös mahdollista luokitella aiheita laajempiin yläkategorioihin ihmisen tekemän tulkinnan mukaan. Esimerkiksi aiheet ”kapitalismi ja sosialismi” sekä ”sosialismi” voitaisiin myös yhdistää saman yläotsikon (’metakoodin’) alle, jolloin yhdistelmäaihe mahdollisesti kuvaisi tarkemmin ristiriitaulottuvuutta. Samoin kansainvälisyyden ja globalisaation aiheet voisi olla mielekästä yhdistää. Tällä tavalla tutkijan tulkinnat tulevat jälleen osaksi aineistoa, mutta voivat selkeyttää aihemallinnuksen tulosten tulkintaa, jos aihemäärä on erityisen suuri.

Avoimena kysymyksenä nouseekin siis, miten aihemallinnuksen soveltajaa voitaisiin paremmin tukea tässä analyyttisessä vaiheessa. Tässä artikkelissa pyrittiin aiheiden valintoja perustelemaan sekä olemassa olevan aihepiiriä käsittelevän kirjallisuuden (Mickelsson 2015) sekä politiikan tutkimuksen teorian kannalta (Kirchheimer 1990; Krouwel 2003; Maas 2001). Aihepiirin kirjallisuutta käytettiin hyödyksi sekä aihemallinnuksen historiallisten muutosten tulkinnassa että myös yksittäisten aiheiden valinnassa – esimerkiksi luonnonlaki-puolueen tunnistaminen perustui kirjallisuuteen tutustumiseen. Toisaalta, laajempi teoriakatsanto mahdollisti aihemallinnuksen käyttämisen lopullisen tuloksen sijasta välivaiheena, jota jatko-operationalisoitiin ja käytettiin teorian tilastollisessa tarkastelussa. Tällöin tehdyille valinnoille voidaan hakea tukea jo olemassa olevasta kirjallisuudesta.

Toisaalta aina ei ole olemassa yhtä vahvaa kirjallisuutta, jonka päälle analyysin voisi rakentaa. Tämän takia viimeaikaisessa kirjallisuudessa on laajasti pohdittu miten laskennallinen analyysi voisi tukeutua ja täydentää perinteistä laadullista työtä triangulaation hengessä (Laaksonen ym. 2017; Muller ym. 2016; Nelson 2017). Kolmas vaihtoehto olisi nähdä aihemallinnus (sekä vastaavat ohjaamattoman koneoppimisen menetelmät) eräänlaisena *grounded theory* -vaiheena. Tämä onkin mielekästä, jos aiheesta ei ole etukäteen tiedossa mahdollisia luokkia ja niiden tulkintoja, vaan kyseessä on avoimempi tutkimussuunta. Suomenkielisinä esimerkkeinä tämän kaltaisesta työstä voi nähdä viimeaikaiset julkaisut agendan hallinnasta sekä kehyksistä (Laaksonen ja Nelimarkka 2018; Ylä-Anttila ym. 2018). Näissä töissä ei kuitenkaan noudateta perinteisen *grounded theory* -menetelmän henkeä esimerkiksi uudelleenkoodauksen tai muistiinpanojen (*memoing*) osalta (vrt. Strauss ja Corbin 1990). Ohjaamattomat

tekstianalyysimenetelmät ovat valitettavasti vielä toistaiseksi avoimia, eikä yleisesti hyväksyttyjä hyviä käytänteitä aiheiden tunnistamiseen, nimeämiseen ja ryhmittelyyn ole olemassa, joten tähän vaiheeseen on tarpeen kiinnittää erityistä huomiota.

Yllä on esitelty kolme erilaista näkökulmaa aiheiden tulkintaan ja sen soveltamiseen, jotka voidaan tiivistää seuraavaksi suositukseksi:

Suositus 2 Aiheiden tulkintaan ja niiden käyttöön voidaan hyödyntää jotain seuraavista kolmesta lähestymistavasta:

- käyttäen olemassa olevaa kirjallisuutta joko sisällöllisenä apuvälineenä tai käyttäen aihemallinnusta välineenä olemassa olevan teorian arvioimiseen
- soveltaen aihemallinnusta sekä aineiston muita analyyseja trianguloiden toistensa kanssa rinnakkain
- kuvaten aihemallinnuksen tuloksissa tehtyjä uudelleenryhmittelyjä sekä kirjaamalla omia havaintoja ja tulkintoja systemaattisesti, kuten aineistolähtöisessä *grounded theory* -prosessissa on ohjeistettu.

Näistä ensimmäinen suositus vastaa ensimmäisessä osatutkimuksessa tunnistettua aiheiden määrän valinnan strategiaa. Toisessa kahdesta tunnistetusta strategiasta määrän valintaan perustui olemassa olevan ennakkokäsityksen hyödyntämiseen mallin valinnassa. Tämän strategian kohdalla onkin syytä olla varovainen, ettei tutkimuksen aikana ensin perustella mallin valintaa olevalla teorialla ja tämän jälkeen perustella mallin mielekkyyttä mallin sopivuudella teoriaan.

KESKUSTELU JA JOHTOPÄÄTÖKSET

Viime aikoina yhteiskuntatieteissä on herännyt mielenkiintoa käyttää ohjaamattomia koneoppimismenetelmiä, kuten aihemallinnusta. Uusien menetelmien avulla tutkimuksen lähestymistavat voivat olla datavetoisia, eksploratiivisia, iteratiivisia sekä suuriin aineistomääriin sopivia (Kitchin 2014). On kuitenkin ilmiselvää, ettei (ohjaamattomien) koneoppimismenetelmien käytön tulisi olla ”teoriatonta” tai ”historiatonta”. Massadatasta (*big data*) ei voida sanoa mitään ilman ymmärrystä kontekstista sekä teorioista (Boyd ja Crawford 2012; Frické 2015). Yhteiskuntatieteilijöiden perinteinen integroituminen alansa oppihistoriaan ja kirjallisuuteen voikin mahdollistaa laadukkaan yhteiskuntatieteellisen tutkimuksen myös (ohjaamattomia) koneoppimismenetelmiä käytettäessä (Grimmer 2015; Wallach 2018).

Tämä tutkimus osaltaan ilmentää, miten ohjaamattomien koneoppimismenetelmien sovelluksissa voidaan sitoutua alan kirjallisuuteen. Sitoutuminen Suomen puoluejärjestelmän historiaan tuki joidenkin aiheiden nimeämistä, esimerkiksi Luonnonlaki-puolueen kohdalla: sanat kuten ’tietoisuus’ sekä ’luonnonlaki’ oli yhdistettävissä tähän puolueeseen. Osatutkimuksessa 2 ehdotetaankin, että aiheiden tulkinnassa ja niiden arvioimisessa tilastollisten menetelmien, kuten aiheiden sisäisen yhteneväisyyden mittaamisen (Chang ym. 2009; Towne ym. 2016), tukena voi myös käyttää teoriaan tai triangulaatioon perustuvia lähestymistapoja. Osatutkimus 2

perustui olemassa olevan kirjallisuuden hyödyntämiseen tulkinnassa, mutta mahdollisuuksia on myös etnografian sekä muiden laadullisten menetelmien käytössä ja yhdistämisessä massadataan (Laaksonen ym. 2017; Muller ym. 2016).

Puolueohjelmien muutosta sekä massapuolueiden kehittymistä kuvaavan tutkimuksen sijaan artikkelin motivoivina tutkimuskysymyksinä oli ymmärtää aihemallinnuksen metodologisia haasteita yhteiskuntatieteille. Kuten osatutkimus 1 toi esille, yhteiskuntatieteellinen tutkimus on perinteisesti suosinut tulkinnallisuutta korostavia lähestymistapoja, joissa tutkija arvioi mikä aihemäärä on helpoiten tulkittavissa. Osatutkimus 1 tarkasteli tutkijoiden lähestymistapoja määrittää tulkinnallisuuden kannalta paras aihemäärä. Tutkimuksessa havaitaan eroja määrissä: neljä eri osallistujaa ehdottivat parhaimmiksi aihemääriksi 25:tä, 28:aa tai 34:ää. Eroja oli myös tulkittavuuden ymmärtämisessä: toiset tutkijat korostivat eri näkökulmien huomioimista ja mukaan tuomista, kun taas toiset pyrkivät yksinkertaistamaan ja vähentämään aiheiden määrää. Osatutkimus 2 näytti, että aiheiden määrällä on vaikutusta analyysin tuloksiin.

Mitä nämä kaksi osatutkimusta kertovat laajemmin ohjaamattomista koneoppimismenetelmistä ja niiden soveltumisesta yhteiskuntatieteelliseen tutkimukseen? Ensin on syytä ymmärtää minkälaisia lupauksia nämä menetelmät ja laskennallinen yhteiskuntatiede on tarjonnut. Laskennallista yhteiskuntatiedettä esittelevässä artikkelissaan Lazer ym. (2009) kutsuvat laskennallisia menetelmiä uudeksi mikroskoopiksi. Heidän mukaan laskennallisten menetelmien avulla voidaan tarkastella yhteiskuntaa uusilla tavoilla. Minkälainen mikroskooppi on siis kyseessä? Kuten Giere (2010) huomauttaa – käyttäen linssejä esimerkkeinä – tieteellinen tieto ei ole koskaan objektiivista, vaan värittynyt käytettyjen instrumenttien kautta. Ohjaamattomat menetelmät ovat haastavia instrumentteja, koska mittavälineen soveltuvuus ongelmaan selviää usein vasta kun analyysi on tehty. Esimerkiksi aihemallinnusprosessin keskeinen vaihe, ohjaamattoman koneoppimisen kautta syntyneiden aiheiden nimeäminen ja tulkinta, on mahdollista vasta kun ne on jo tuotettu. Sitä ennen on täysin tuntematonta, minkä kaltaisia ryhmiä aineistosta ”nousee”. Nämä menetelmät ovat usein myös satunnaisuutta käyttäviä, jolloin saman menetelmän käyttäminen ei välttämättä johda sellaisenaan samoihin tuloksiin. Symons ja Alvarado (2016) argumentoivatkin, että laskennallisen analyysin virhelähteisiin tulisi keskittyä aiempaa enemmän tulosten tulkinnass

Tarkoittaako yllä esitetty laskennallisten menetelmien kritiikki sitä, ettei tieteellisessä yhteisössä voida ajatella ohjaamattomien koneoppmismenetelmien tuottavan tieteellistä tietoa? Tilanne ei ole näin vakava. Esimerkiksi Giere (2010) keskittyy kritiikissään siihen, että tieteellinen havainnointi on mahdollista kuvata läpinäkyvästi. Tämä pätee myös laadulliseen tutkimukseen, missä on tarpeen reflektoida myös tutkijan roolia osana tulkintaa ja tuloksia. Tällä hetkellä tämä ei kuitenkaan ole yleinen tapa ohjaamattoman koneoppimisen käyttämisessä (ks. loppuviite 8). Henkilön omalla taustalla ja kokemuksella on merkitystä esimerkiksi teoriaviitekehysten valinnassa sekä laajemmin tulkinnassa. Esimerkiksi osatutkimuksessa 1 havaitut erot osaltaan heijastelivat eri koehenkilöiden taustojen merkitystä. Tämän takia ohjaamattoman koneoppimisen sovelluksiin voitaisiin pyrkiä soveltamaan samanlaisia käytänteitä kuin laadullisessa tutkimuksessa. Tämän tavoite on nostaa esille myös ohjaamattomassa koneoppimista soveltavien tutkimusprojektien useat erilaiset valinnat.

Suositus 3 Tutkimustyön raportoinnin tulisi olla refleksiivisempää. Siinä olisi kuvattava tutkijoiden taustoja ja pohtia mahdollisia syitä tulkintoihin. Lisäksi tutkimustyön aikana tehtyjä ei-julkaistuja analyysejä, tehtyjä valintoja sekä rajauksia on tarpeen esitellä.

Samaan aikaan kun laskennallinen yhteiskuntatiede suosittaa algoritmien käyttöä tutkimusongelmien ratkaisemiseen, kriittisen algoritmitutkimuksen koulukunta tuo esille algoritmisten järjestelmien yhteiskunnallisia vaikutuksia (esim. Kitchin 2017; Gillespie 2012). Kriittisen algoritmitutkimuksen sanoma on – tiivistäen – että algoritmien kehittäjillä ja suunnittelijoilla on näkymätöntä valtaa (esim. Beer 2017). Esimerkiksi algoritmien aiheuttama syrjintä päätöksentekotilanteessa on ollut aktiivinen tutkimusalue (esim. Burrell 2016). Hiljaittain Es ym. (2018) ovat pyrkineet nostamaan samaa ongelmanasettelua esille myös tieteellisen tutkimuksen piirissä kritisoidessaan tutkimusta varten luotuja työkaluja ja vaatiessaan niiden tarkempaa tutkimusta.

Kritiikki on erittäin osuvaa myös laskennallisen yhteiskuntatieteen suhteen. Onko mahdollista, että laskennallisessa tutkimuksessa käytettävissä koneoppimisalgoritmeissä on samanlaisia vallankäyttöön liittyviä piirteitä? Johtavatko tietyt tekstianalyysin menetelmät systemaattisiin väärintulkintoihin? Esimerkiksi sanojen irrottaminen lausejärjestyksestä (*bag of words*) voi johtaa virhetulkintoihin. Toisaalta, onko syytä suosittaa nimenomaisesti tiettyä mittaria aihemäärän valintaan, kuten tässä työssä on tehty?

Yllä esitetyt kysymykset ovat avoimia, mutta tieteellisen tutkimustyön kannalta keskeisiä. Huolimaton tai virheellinen ohjaamattoman menetelmän käyttö saattaa 'pakottaa' menetelmän tuottamaan tiettyjä tuloksia – ja nämä tulokset voivat luoda vääristyneen näkökulman maailmaan. Tämän takia osatutkimuksen 2 pohdinnat tulosten tulkinnasta korostivat tarvetta pohtia löydöksiä suhteessa muihin menetelmiin ja teorioihin. Toistaiseksi kriittisen algoritmitutkimuksen kirjallisuus ei ole käsitellyt mitään tiettyjä menetelmiä tai lähestymistapoja tarkasti tieteellisen tutkimustyön osana. Toistaiseksi voidaankin antaa vain seuraava yleinen suositus:

Suositus 4 Laskennallisen yhteiskuntatieteen soveltajien tulee seurata menetelmäkehityskeskustelun lisäksi kriittisen algoritimikirjallisuuden havaintoja esimerkiksi algoritmien puolueellisuudesta.

Artikkeli on laajentanut suomalaista keskustelua tekstiaineistojen automaattisen analysoinnin mahdollisuuksista. Artikkelissa on kyseenalaistettu tyypillisimpiä yhteiskuntatieteissä hyödynnettyjä lähestymistapoja aihemallinnuksen käyttöön. Vastaava kritiikki voidaan laajentaa koskemaan myös muita ohjaamattomia menetelmiä. Artikkelissa esitetty keskustelu tutkimusprosessiin liittyvistä valinnoista sekä toteutettu esimerkinomainen sovellus osoittavatkin, etteivät ohjaamattomat menetelmät ole automatisoitu prosessi, vaan vaatii aina tutkijan omaa tulkintaa ja päätöksentekoa.

Tutkimuksen ensimmäisessä osatutkimuksessa havaittiin, että tulkinnallisuutta korostavassa aihemallinnustyössä aihemäärät vaihtelevat runsaasti. Toisessa osatutkimuksessa taas näytettiin, että aihemäärällä oli vähäisiä, mutta merkillepantavia eroja empiirisissä tuloksissa. Yhdessä osatutkimukset nostavat esille aihemallinnuksen luotettavuuteen ja toistettavuuteen liittyviä haastavia kysymyksiä. Kysymysten ratkaisemiseksi artikkelissa suositeltiin, että aihemallinnuksessa

siirryttäisiin käyttämään tilastollista aiheiden määrän valintaa sekä muodostettaisiin selkeä yhteys olemassa olevaan yhteiskuntatieteelliseen teoriaan. Lopulta on syytä kuitenkin huomata, laskennallisia menetelmiä kohtaan voi syystä olla kriittinen. Tutkimusprosessin läpinäkyvyyteen ja tehtyjen valintojen perusteluihin tulee vaatia tarkempaa raportointia.

KIITOKSET

Matti Nelimarkka kiittää Koneen Säätiötä sekä Suomen Akatemiaa tutkimukseen saadusta rahoituksesta. Lisäksi Matti Nelimarkka kiittää useita artikkeliin saatuja kommentteja Rajapinta-yhdistyksen tapahtumissa, Digital Content Communities-tutkimusryhmältä sekä artikkelin vertaisarvioijilta.

VIITTEET

1. Aihemallinnuksessa on kolme hyperparametirä: α ja β säätelevät aiheiden ja sanojen sekä aiheiden ja dokumenttien todennäköisyysjakaumia ja k aiheiden lukumäärää.
2. http://www.nltk.org/api/nltk.stem.html
3. http://www.kielipankki.fi/tyokalut/
4. Tämän artikkelintulosten pohjalta voidaan kyseenalaistaa yleisesti tehtyä tulkintaa välttää tilastollisia menetelmia aihemäärän valinnassa. Myös ihmisten tulkinnassa voidaan selkeästi päätyä tilanteeseen, jossa eri osallistujien mielestä selkein tulkinta on varsin erilainen.
5. POHTIVA-tietokanta: http://www.fsd.uta.fi/pohtiva/https://www.fsd.uta.fi/pohtiva/.
6. Tietotekniikan tutkimuksessa asetelmia, missä henkilö käyttää ohjelmistoa tutkimusta varten, kutsutaan käyttäjäkokeiksi (*user study*). Tätä ei pidä sekoittaa yhteiskuntatieteessä yleisemmin käytettyyn satunnaiseen koeasetelmaan (*randomized controlled trial*) tai luonnolliseen kokeeseen (*natural experiment*).
7. Visualisaatio-ohjelma on saatavissa osoitteesta https://github.com/HIIT/topicmodel-viz.
8. Aihemallinnuksen tutkimustraditiossa ei ole toistaiseksi tapana tuoda esille tämän kaltaista subjektiivisuutta tai laajemmin reflektoida tutkijan omaa asemaa analyysiprosessissa. Noudattaen tätä toimintatapaa, tässä artikkelissa ei käsittellä osallistujien ennakkokäsityksiä tai taustaa.
9. Mukaillen Dragicevic ym. (2014) artikkelia, missä näytetään p-arvon heikkous toistamalla samaa koeasetelmaa ja näyttämällä erilaisia tuloksia, voitaisiin tästäkin kenties kirjoittaa viisi eri tieteellistä artikkelia ja jokaisessa keskustella juuri kyseisen mallin tuottamista erityispiirteistä.
10. Kymmenen prosentin raja-arvo on varsin mielivaltainen: jollekin toiselle merkittävyyden raja voisi olla 5%, 7.5% tai 12.5%. Ajatuksena taustalla on, että jos puolueohjelma ottaa kantaa johonkin aiheeseen vähintään tai yli kymmenyksellä koko tekstiaineistosta, on tämä aihe puolueohjelmassa merkittävästi esillä.
11. Eräs ratkaisu olisikin tarkastella liikkeiden ja puolueiden institutionalisoitumisprosessia, eli arvioida kuinka kunkin puolueen merkittävien aiheiden määrä on muuttunut vuosien varrella. Tässäkin lähestymistavassa olisi kuitenkin ollut ilmeisiä haasteita: Puolueiden välillä voi olla eroja niiden

kehitysnopeudessa kohti yleispuolueita. Toisaalta puolue itsessään on jo institutionalisoitunut ja analyysin olisi vaikea huomioida muita puolueiden toimintaan vaikuttavia tekijöitä, kuten teknologian kehitystä (vrt. Farrell ja Webb 2000). Argumentaation selkeyden kannalta on tarkoituksenmukaista rajata tarkastelu ajanjaksoon ennen uusien liikkeiden syntymistä ja arvioida puolueohjelmia tämän aikakauden yhteydessä.

12. Wilcoxonin testi on parametriton testi yhteiskuntatieteilijöille tutummasta t-testistä, jota käytetään artikkelissa, koska ei ole syytä olettaa aineiston olevan normaalijakautunut. Koska tässä on tehty useita p-testauksia, taulukossa esitettävät luvut on korjattu Bernoulli-kertoimella.

LÄHTEET

Aletras, Nikolaos, Baldwin, Timothy, Lau, Jey Han ja Stevenson Mark. 2017. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology* 68:1, 154–67.

Arter, David. 1999. From class party to catchall party?: The adaptation of the Finnish Agrarian-Center Party. *Scandinavian Political Studies* 22:2, 157–80.

Bandalos, Deborah ja Meggen Boehm-Kaufman. 2010. "Four Common Misconceptions in Explanatory Factor Analysis." In *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in Organizational and Social Sciences*, eds. Charles E Lance, Charles E Lance, and Robert J Vandenberg. Routledge.

Beer, David. 2017. The social power of algorithms. *Information, Communication ja Society* 20:1, 1–13.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55:4, 77–84.

Blei, David M., Ng, Andrew Y. ja Jordan, Michael I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3, 993–1022.

Boyd, Danah ja Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication ja Society* 15:5, 662–79.

Burrell, Jenna. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data ja Society* 3:1, 1–12.

Chang, Jonathan, Gerrish, Sean, Wang, Chong ja Blei David M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems* 22, 288–296.

Cioffi-Revilla, Claudio. 2010. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics* 2:3, 259–71.

Denny, Matthew James ja Spirling Arthur. 2018. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do About It. *Political Analysis* 26:2, 168–89.

Es, Karin van, Wieringa, Maranke ja Schäfer Mirko Tobias. 2018. Tool Criticism: From Digital Methods to Digital Methodology. *Proceedings of the 2nd International Conference on Web Studies - Ws.2 2018*. New York: ACM Press, 24–27.

Fabrigar, Leandre R., Wegener, Duane T., MacCallum, Robert C. ja Strahan Erin J. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4:3, 272–99.

Farrell, David M. ja Paul Webb. 2000. Political Parties as Campaign Organizations. Teoksessa Russel J. Dalton ja Martin P. Wattenberg (toim.) *Parties Without Partisans: Political Change in Advanced Industrial Democracies*. Oxford: Oxford University Press, 102–28.

Frické, Martin. 2015. Big data and its epistemology. *Journal of the Association for Information Science and Technology* 66:4, 651–61.

Giere, Ronald N. 2010. *Scientific Perspectivism*. Chicago: University of Chicago Press.

Gillespie, Tarleton. 2012. The relevance of algorithms. Teoksessa Tarleton Gillespie ym. (toim.), *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge: MIT Press, 167–94.

Greene, Derek, O'Callaghan, Derek ja Cunningham Pádraig. 2014. How many topics? Stability analysis for topic models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 8724 LNAI(PART 1), 498–513.

Griffiths, Thomas L. ja Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(Supplement 1), 5228–35.

Grimmer, Justin. 2015. We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science ja Politics* 48:1, 80–83.

Grimmer, Justin ja Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21:3, 267–97.

Jacobi, Carina, van Atteveldt, Wouter ja Welbers Kasper. 2016. Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling. *Digital Journalism* 4:1, 89–106.

Jurek, Steven J. ja Anthony Scime. 2014. Achieving Democratic Leadership: A Data-Mined Prescription. *Social Science Quarterly* 95:1, 97–110.

Kirchheimer, Otto. 1990. The Catch-All Party. *The West European Party System*: 50–60.

Kitchin, Rob. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data ja Society* 1:1, 1–12.

Kitchin, Rob. 2017. Thinking critically about and researching algorithms. *Information, Communication ja Society* 20:1, 14–29.

Krouwel, André. 2003. Otto Kirchheimer and the catch-all party. *West European Politics* 26:2, 23–40.

Laaksonen, Salla-Maaria Maaria, Nelimarkka, Matti, Tuokko, Mari, Marttila, Mari, Kekkonen, Arto ja Villi Mikko. 2017. Working the fields of big data: Using big-data-augmented online ethnography to study candidate–candidate interaction at election time. *Journal of Information Technology and Politics* 14:1, 110–31.

Laaksonen, Salla-Maaria ja Nelimarkka Matti. 2018. Omat ja muiden aiheet : Laskennallinen analyysi vaalijulkisuuden teemoista ja aiheomistajuudesta. *Politiikka* 60:2, 132–47.

Lazer, David ym. 2009. Social science. Computational social science. *Science* 6:323, 721–23.

Levy, Karen E. C. ja Franklin Michael. 2014. Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking. *Social Science Computer Review* 32:2, 182–94.

Maas, Willem. 2001. Catch-all parties. Teoksessa Jonathan Michie (toim.) *Reader's Guide to the Social Sciences*. Abingdon: Routledge, 167–168.

Metsämuuronen, Jari. 2003. Tutkimuksen Tekemisen Perusteet Ihmistieteissä. Helsinki: International Methelp.

Mickelsson, Rauli. 2015. *Suomen Puolueet: Vapauden Ajasta Maailmantuskaan*. Tampere: Vastapaino.

Mohr, John W. ja Bogdanov Petko. 2013. Introduction-Topic models: What they are and why they matter. *Poetics* 41:6, 545–69.

Muller, Guhab, Shion, Michael, Baumerc, Eric P.S., Mimnoc, David ja Shami N. Sadat. 2016. Machine Learning and Grounded Theory Method. *Proceedings of the 19th International Conference on Supporting Group Work - Group '16*. New York: ACM Press, 3–8.

Nelimarkka, Matti, Laaksonen, Salla-Maaria, Marttila, Mari, Kekkonen, Arto, Tuokko, Mari ja Villi

Mikko. Influencing the News Through Social Media: Online Agenda Building and Normalization During a Pre-Electoral Campaign Period. Julkaisematon konferenssiesitys. *66th Annual International Communication Association (ICA) Conference*. Fukuoka, Japani, Kesäkuu 2016.

Nelson, Laura K. 2017. Computational Grounded Theory. *Sociological Methods ja Research*.

Purhonen, Semi ja Arho Toikka. 2016. 'Big datan' haaste ja uudet laskennalliset teksiaineistojen analyysimenetelmät. *Sosiologia* 53:1, 6–26.

Russell, Daniel W. 2002. In Search of Underlying Dimensions: The Use (and Abuse) of Factor Analysis in Personality and Social Psychology Bulletin." *Personality and Social Psychology Bulletin* 28(12): 1629–46.

Savage, Mike. 2013. The 'Social Life of Methods': A Critical Introduction. *Theory, Culture ja Society* 30:4, 3–21.

Scarrow, Susan E. 2000. Parties Without Members? Party Organization in a Changing Electoral Environment. Teoksessa Russel J. Dalton ja Martin P. Wattenberg (toim.) *Parties Without Partisans: Political Change in Advanced Industrial Democracies*. Oxford: Oxford University Press, 102–28.

Schofield, Alexandra ja Mimno David. 2016. Comparing Apples to Apple : The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics* 4, 287–300.

Strauss, Anselm ja Corbin Juliet M. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques.* Sage Publications.

Symons, John ja Alvarado Ramón. 2016. Can we trust Big Data? Applying philosophy of science to software. *Big Data ja Society* 3:2.

Towne, W. Ben, Rosé, Carolyn P. ja Herbsleb James. 2016. Measuring Similarity Similarly: LDA and Human Perception. *ACM Transactions on Intelligent Systems and Technology ACM Reference Format ACM Trans. Intell. Syst. Technol* 7:2, 25:1–25:29.

Wallach, Hanna. 2018. Computational social science ≠ computer science + social data. *Communications of the ACM* 61:3, 42–44.

Wallach, Hanna M., Mimno, David ja Mccallum Andrew. 2009. Rethinking LDA: Why Priors Matter. *Advances in Neural Information Processing Systems* 22:2, 1973–81.

Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan ja Mimno David. 2009. Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning - Icml '09* New York: ACM Press, 1–8.

Watts, Duncan J. 2011. *Everything Is Obvious:* Once You Know the Answer*. New York: Crown Business.

Ylä-Anttila, Tuukka, Eranti, Veikko ja Kukkonen Anna. 2018. Aihemallinnuksesta Kehitysmallinnukseen. *Politiikka* 60:2, 158–56.

Yu, Bei, Kaufmann, Stefan ja Diermeier Daniel. 2008. Classifying Party Affiliation from Political Speech. *Journal of Information Technology ja Politics* 5:1, 33–48.

KIRJOITTAJATIEDOT

MATTI NELIMARKKA
FT, VTM, yliopistonlehtori
Valtiotieteellinen tiedekunta
Helsingin yliopisto
Tietotekniikan laitos
Aalto-yliopisto
matti.nelimarkka@helsinki.fi

	Tulkinta	$k = 25$	$k = 28$	$k = 33$	$k = 34$
Suuret kysymykset					
	Kapitalismi ja sosialismi	kapitalismi, kapitalisti, kehitys, kommunisti, puolue, sosialismi, sosialistinen, taistelu, työväenluokka, yhteiskunta	kapitalismi, kapitalisti, kehitys, kommunisti, maailma, politiikka, sosialismi, sosialistinen, yhteiskunnallinen, yhteiskunta	kapitalismi, kapitalisti, kehitys, kommunisti, maailma, politiikka, pääoma, sosialismi, sosialistinen, yhteiskunta	kapitalismi, kapitalisti, kommunisti, kommunistinen, maailma, politiikka, pääoma, sosialistinen, yhteiskunnallinen, yhteiskunta
	Isänmaa	henki, isänmaa, johtaa, kansa, kansallinen, perusajatus, pääoma, suomalainen, suomi, suomia	henki, isänmaa, kansa, kansallinen, perusajatus, pääoma, suomalainen, suomi, suomia, tarkoittaa	henki, isänmaa, kansa, kansallinen, perusajatus, suomalainen, suomi, suomia, tarkoittaa, toimia	henki, isänmaa, kansa, maa, perusajatus, pääoma, suomalainen, suomi, suomia, toimia
	Kansalaisuus	kansa, kansalainen, mahdollisuus, saada, taloudellinen, toiminta, valtio, vapaus, yhteiskunta, yksityinen	kansa, kansalainen, mahdollisuus, sosiaalinen, taloudellinen, toiminta, vapaa, vapaus, yhteiskunta, yksilö	ihminen, kehitys, markkinat, poliittinen, sosiaalinen, tuotanto, vapaus, vasemmisto, yhteiskunta, yritys	kansa, kansalainen, pyrkiä, saada, taloudellinen, tehtävä, toiminta, turvata, vapaus, yhteiskunta
Puolueaiheet					
	Keskusta	henkinen, ihminen, kehitys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomi, turvata	henkinen, ihminen, kehitys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomalainen, suomi	ihminen, ihmisyys, kehitys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomalainen, tulevaisuus	aineellinen, henkinen, ihminen, ihmisyys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomalainen
	Maalaisliitto				elämä, kansa, kansallinen, lainsäädäntö, maa, maalaisliitto, maaseutu, olo, sivistys, taloudellinen
	Maalaisliitto			kansa, maa, maalaisliitto, maaseutu, maatalous, olo, sivistys, talous, työ, valtio	alue, elintaso, kuivatus, kuulua, laitos, maalaisliitto, maaseutu, pienteollisuus, prosentti, vuosikymmen
	Luonnonlaki	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, ryhmä, tietoisuus, tutkimus, yhteiskunta	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, taso, tietoisuus, tutkimus, yhteiskunta	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, taso, tietoisuus, tutkimus, yhteiskunta	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, taso, tietoisuus, tutkimus, yhteiskunta
	Kristilliset	demokraatti, hoito, koulu, kristillinen, lapsi, ohjelma, perhe, päihde, tarve, tuki	demokraatti, hoito, kristillinen, käyttö, lapsi, nuori, parantaa, perhe, saada, tuki	demokraatti, elämä, hoito, kristillinen, lapsi, nuori, perhe, päihde, saada, tuki	demokraatti, euro, hoito, koulu, kristillinen, lapsi, ohjelma, perhe, päihde, tuki
	Piraattipuolue	demokratia, kansalainen, muutos, oikeus, piraatti, puolue, sananvapaus, suomi, tieto, viestintä		demokratia, kansalainen, käyttö, muutos, oikeus, piraatti, puolue, sananvapaus, tieto, vapaus	demokratia, kansalainen, käyttö, oikeus, piraatti, poista, puolue, sananvapaus, tieto, viestintä
	Vihreät	demokratia, eläin, esimerkki, maailma, oikeus, perus, reilu, tukea, unioni, vihreä	energia, kulutus, käyttö, luonto, päästö, suomi, tukea, vero, vihreä, ympäristö	energia, esimerkki, käyttö, luonto, nosta, päästö, suomi, vero, vihreä, ympäristö	energia, kulutus, käyttö, luonto, päästö, suomi, tukea, vero, vihreä, ympäristö
	Vihreä unioni	asunto, energia, eurooppa, liitto, luku, luonto, nainen, ongelma, suomi, ympäristö	avoimuus, demokratia, esimerkki, nykyinen, oikeus, reilu, tieto, tulo, unioni, vihreä	demokratia, eläin, esimerkki, globalisaatio, kansalaisuus, nykyinen, perus, reilu, unioni, vihreä	demokratia, eläin, esimerkki, oikeudenmukaisuus, perus, reilu, tukea, tulo, unioni, vihreä
	Sosialismi	kapitalismi, kapitalisti, kehitys, kommunisti, puolue, sosialismi, sosialistinen, taistelu, työväenluokka, yhteiskunta	kansa, kapitalistinen, kehitys, maa, puolue, sosialismi, sosialistinen, suomi, tuotanto, voima	etu, kansa, kapitalistinen, maa, puolue, sosialismi, sosialistinen, suomi, tuotanto, voima	kansa, kapitalistinen, kehitys, maa, puolue, sosialismi, sosialistinen, suomi, tuotanto, voima
	Kokoomus			julkinen, kokoomus, luku, markkinatalous, osake, rahoitus, sosiaalinen, teollisuus, vero, yritys	elinkeinoelämä, julkinen, kansalainen, kokoomus, luku, markkinatalous, osake, valta, vero, yritys
	Kokoomus			eettinen, henkinen, kilpailu, kokoomus, moniarvoisuus, valinta, vastuu, yksilö, yksilöllinen, yksilöllisyys	henkinen, kansanvalta, kilpailu, kokoomus, toimenpide, turvallisuus, valinta, yksilö, yksilöllinen, ympäristönsuojelu
	Työväenpuolue	eurooppa, hyvinvointi, kansa, pääoma, suomi, tavoite, työväenpuolue, unioni, vaihtoehto, yhteinen	eurooppa, hyvinvointi, kansa, kunta, pääoma, suomi, tavoite, toiminta, työväenpuolue, yhteinen	hyvinvointi, järjestö, kansa, katsoa, sota, suomi, tavoite, työväenpuolue, yhteinen, yhteiskuntapolitiikka	hyvinvointi, järjestö, kansa, katsoa, nykyinen, sota, suomi, tavoite, työväenpuolue, yhteinen
	Vasemmistoliitto	arvo, kehitys, liitto, markkinat, oikeus, poliittinen, sosiaalinen, vapaus, vasemmisto, yritys	ehdottaa, hyvinvointivaltio, kehitys, liitto, nainen, ohjelma, oikeus, pääoma, suomi, vasemmisto	ehdottaa, hyvinvointivaltio, liitto, nainen, oikeus, palkka, pääoma, sukupuoli, tasainen, vasemmisto	ehdottaa, hyvinvointivaltio, liitto, nainen, ohjelma, oikeus, pitää, pääoma, tasainen, vasemmisto
	RKP	kansanpuolue, kieli, maa, oikeus, pitää, puolue, ruotsalainen, ruotsinkielinen, suomi, väestö	kansanpuolue, kieli, pitää, puolue, ruotsalainen, ruotsi, ruotsinkielinen, suomenruotsalainen, vaatia, väestö	edistä, kansanpuolue, kieli, puolue, ruotsalainen, ruotsi, ruotsinkielinen, suomenruotsalainen, suomi, väestö	edistä, kansanpuolue, kieli, pitää, puolue, ruotsalainen, ruotsi, ruotsinkielinen, suomenruotsalainen, väestö
	Liberaalit	eläke, julkinen, kansalainen, liberaali, oikeus, palvelu, perus, tulo, työ, yksityinen	ehdokas, eläke, julkinen, kansalainen, laki, liberaali, oikeus, perus, tulo, vapaus	eläke, julkinen, kansalainen, laki, liberaali, perus, tulo, vapaus, vero, verotus	eläke, julkinen, kansalainen, liberaali, palvelu, perus, pitää, tehtävä, tulo, vakuutus
	SDP		demokratia, julkinen, kansalainen, sosialidemokraattinen, taloudellinen, tavoite, toteuttaminen, työntekijä, valta, yhteiskunnallinen	muutos, parantaminen, sdp, sosialidemokraattinen, taloudellinen, tasainen, toteuttaminen, työntekijä, työväenliike, uudistaminen	kansanvalta, puolue, sosialidemokraattinen, tavoite, toteuttaminen, työ, työntekijä, työväenliike, valta, yhteiskunnallinen
	SKDL			alista, demokraattinen, kehittäminen, kehitys, liike, skdl, toiminta, työntekijä, vaihtoehto, yhteistyö	demokraattinen, henkinen, laaja, liitto, luoda, maa, poliittinen, skdl, yhteiskunnallinen, yhteistyö
	Perussuomalaiset			kansa, kansalainen, maa, perus, poista, puolue, saada, suomalainen, suomi, turvata	etu, kansa, kansalainen, maa, perus, puolue, saada, suomalainen, suomi, turvata
Politiikka-aiheet					
	Yrittäjyys		kansalainen, kilpailu, koti, nuori, sosiaalinen, teollisuus, tuotanto, valtiovalta, yrittäjä, yritys	edellytys, kansainvälinen, kansalainen, kehittäminen, kehitys, kulttuuri, taloudellinen, toiminta, yhteiskunnallinen, yhteiskunta	kehitys, markkinat, oikeus, poliittinen, sosiaalinen, tuotanto, vapaus, vasemmisto, yhteiskunta, yritys
	Inhimillisyys		huomio, ihminen, ihmisyys, inhimillinen, maailma, mahdollinen, pitää, puolue, vaatia, yhteiskunnallinen	huomio, ihminen, ihmisyys, inhimillinen, maailma, mahdollinen, pitää, puolue, tekijä, vaatia	huomio, ihminen, ihmisyys, inhimillinen, maailma, mahdollinen, oikeuslaitos, puolue, taide, vaatia
	Demokratia	demokraattinen, demokratia, kansa, kehitys, maa, suomi, taloudellinen, tuotanto, yhteiskunnallinen, yhteistyö	demokratia, ihminen, kehitys, markkinat, oikeus, poliittinen, sosiaalinen, vapaus, vasemmisto, yhteiskunta		
	Alueet	ala, alue, asunto, laina, luku, myöntää, perhe, suorittaa, teollisuus, toimenpide	alue, hinta, kaupunki, korko, laina, laitos, maatalous, palkka, prosentti, pula	alue, hinta, kaupunki, laitos, piiri, pohjoinen, seutu, teollisuus, toiminta, vesi	
	Luonto ja ympäristö	alue, energia, ihminen, käyttö, lapsi, liikenne, luonto, saada, suomi, ympäristö	ihminen, jäte, koulu, käyttö, lapsi, metsä, saada, suomi, tavoite, ympäristö	ihminen, kulttuuri, käyttö, lapsi, liikenne, luonto, metsä, suomi, turku, ympäristö	energia, ihminen, jäte, käyttö, lapsi, metsä, saada, suomi, turku, ympäristö
	Alueet			alueellinen, hallinto, huomio, ihminen, kunnallinen, kunta, mahdollisuus, saattaa, tieto, valtio	
	Edut ja turvaaminen	estä, etu, haluta, ihmisyys, kansa, pitää, puolue, saada, työ, vaatia	etu, kansa, kansalainen, maa, poista, puolue, saada, suomalainen, suomi, turvata		
	Verotus	eurooppa, kehittää, koulutus, palvelu, suomi, taso, tavoite, tuki, työ, verotus	järjestelmä, koulutus, kunta, palvelu, suomi, taso, tuki, työ, vero, yritys		eurooppa, koulutus, kunta, palvelu, suomi, taso, tuki, työ, verotus, yritys
	Kieli ja koulutus	kieli, kunta, laki, maa, määrätä, oikeus, saada, suomi, valtio, yleinen	koulu, kunta, laki, maa, määrätä, oikeus, saada, suomi, valtio, yleinen	kansa, koulu, kunta, laki, maa, oikeus, saada, suomi, valtio, yleinen	kieli, kunta, laki, maa, määrätä, oikeus, saada, suomi, valtio, yleinen
	Mahdollisuudet	maa, nykyinen, osa, pitää, saada, suomi, tehdä, tärkeä, valtio, voida	maa, mahdollinen, nykyinen, osa, saada, suomi, tehdä, työ, valtio, voida	käyttää, maa, nykyinen, osa, pitää, saada, suomi, tehdä, valtio, voida	järjestelmä, maa, nykyinen, ongelma, osa, saada, suomi, tehdä, valtio, voida
	Edut ja turvaaminen	etu, kansa, köyhälistö, pula, suomi, tahtoa, työ, työttömyys, työväki, yhteinen		etu, kansa, korko, köyhälistö, pula, rintama, tahtoa, työ, työväki, yhteinen	etu, korko, köyhälistö, pula, rintama, tahtoa, työ, työväki, valta, yhteinen
	Hyvinvointipalvelut	ehdottaa, hyvinvointi, kunnallinen, kunta, lapsi, oikeus, palvelu, suomalainen, työntekijä, yksityinen		esimerkki, hyvinvointi, kunta, maksu, malli, osaaminen, palvelu, suomi, tuki, unioni	hyvinvointi, kunnallinen, kunta, kuntalainen, malli, palvelu, tarve, terveys, vanhus, yksityinen
	Arvot	arvo, ihminen, kansalainen, kulttuuri, luonto, mahdollisuus, talous, työ, yhteiskunta, ympäristö	arvo, ihminen, kansalainen, kehitys, mahdollisuus, talous, työ, voida, yhteiskunta, ympäristö	arvo, ihminen, kansalainen, mahdollisuus, oikeus, suomi, talous, työ, yhteiskunta, ympäristö	arvo, ihminen, kansalainen, kehitys, mahdollisuus, oikeus, työ, voida, yhteiskunta, ympäristö
	Maaseudun asema	asema, huomio, kansa, maa, maaseutu, maatalous, sivistys, taloudellinen, työ, valtio	huomio, kansa, kehittää, maa, maaseutu, maatalous, taloudellinen, toiminta, työ, valtio	huomio, kansa, kansalainen, maa, mahdollisuus, pyrkiä, toiminta, työ, valtio, yhteiskunta	asema, huomio, maa, maatalous, saada, tarve, toimenpide, toiminta, työ, valtio
	Kehittäminen ja mahdollisuudet	huomio, kehittäminen, kehittää, koulutus, lisätä, mahdollisuus, taloudellinen, toiminta, turvata, työ	huomio, kehittää, koulutus, kunta, lisätä, maa, mahdollisuus, taloudellinen, toiminta, turvata	järjestelmä, kehittää, koulutus, lisätä, maa, mahdollisuus, tavoite, toiminta, turvata, työ	järjestelmä, kehittäminen, kehittää, koulutus, lisätä, mahdollisuus, taloudellinen, tavoite, toiminta, turvata
	Miehet ja naiset		asunto, eurooppa, mies, nainen, neuvosto, puolue, suomalainen, taloudellinen, vuosikymmen, yhteinen	henkinen, korkea, kulttuuri, lama, mies, nainen, puolue, suomalainen, tuoda, vuosikymmen	asunto, eurooppa, keskeinen, kulttuuri, mies, nainen, taloudellinen, vuosikymmen, yhteinen, ympäristö
	Uskonto		elämä, elää, hengellinen, jumala, koti, kristillinen, kristinusko, tehtävä, tuki, yhteisö		
	Ihmiskunta				alista, eurooppa, ihmiskunta, kapitalismi, kasvaa, markkinat, raha, työttömyys, unioni, vaihtoehto